

HANS MARK



ENCYCLOPEDIA OF

SPACE SCIENCE

and

TECHNOLOGY

ENCYCLOPEDIA OF

SPACE SCIENCE
AND
TECHNOLOGY

VOLUME 1

ENCYCLOPEDIA OF SPACE SCIENCE AND TECHNOLOGY

Editor

Hans Mark

The University of Texas at Austin

Associate Editors

Milton A. Silveira

Principal Engineer, Aerospace Corp.
University of Vermont

Michael Yarymovych

President International Academy of
Astronautics

Editorial Board

Vyacheslav M. Balebanov

Russian Academy of Sciences

William F. Ballhaus, Jr.

The Aerospace Corporation

Robert H. Bishop

University of Texas at Austin

Aaron Cohen

Texas A & M University

Wallace T. Fowler

University of Texas at Austin

F. Andrew Gaffney

Vanderbilt University Medical Center

Owen K. Garriott

University of Alabama

Tom Gehrels

University of Arizona at Tucson

Gerry Griffin

GDG Consulting

Milton Halem

NASA-Goddard Space Flight Center

John S. Lewis

University of Arizona at Tucson

Thomas S. Moorman

Booz Allen & Hamilton

Norman F. Ness

University of Delaware

Robert E. Smylie

National Aeronautics Space
Administration

Richard H. Truly

National Renewable Energy Laboratory

Albert D. Wheelon

Hughes Aircraft Co.

Peter G. Wilhelm

U.S. Naval Research Laboratory

Laurence R. Young

Massachusetts Institute of Technology

Alexander Zakharov

Russian Academy of Sciences

Managing Editor

Maureen Salkin

Editorial Staff

Vice President, STM Books: **Janet Bailey**

Executive Editor:

Jacqueline I. Kroschwitz

Director, Book Production and Manufacturing:

Camille P. Carter

Managing Editor: **Shirley Thomas**

Illustrations Manager: **Dean Gonzalez**

Assistant Managing Editor:

Kristen Parrish

Editorial Assistant: **Surlan Murrell**

ENCYCLOPEDIA OF

SPACE SCIENCE —AND— TECHNOLOGY

VOLUME 1

Hans Mark

Editor

Milton Silveira

Associate Editor

Michael I. Yarymovych

Associate Editor

Maureen Salkin

Managing Editor

The *Encyclopedia of Space Science and Technology* is available Online in full color
at www.interscience.wiley.com/esst

 **WILEY-INTERSCIENCE**

A John Wiley & Sons, Inc., Publication

Copyright © 2003 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permreq@wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. This advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in Publication Data:

Encyclopedia of Space Science & Technology/Hans Mark [editor].

p. cm.

Includes index.

ISBN 0-471-32408-6 (set: acid-free paper)

1. Space Science—Encyclopedias. I. Title: Encyclopedia of Space Science and Technology.

II. Mark, Hans, 1929-

QB497.E53 2003

500.5'03—dc21

2002028867

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

High Flight

By Pilot Officer John G. Magee, Jr., RCAF

Oh, I have slipped the surly bonds of earth
And danced the skies on laughter-silvered wings;
Sunward I've climbed, and joined the tumbling mirth
Of Sun-split clouds – and done a hundred things
You have not dreamed of – wheeled and soared and swung
High in the sunlit silence. Hov'ring there.
I've chased the shouting wind along, and flung
My eager craft through footless halls of air.
Up, up the long, delirious, burning blue
I've topped the windswept heights with easy grace
Where never lark, or even eagle flew.
And, while with silent, lifting mind I've trod
The high untrespassed sanctity of space
Put out my hand, and touched the face of God.

Pilot Officer John Gillespie Magee, Jr., an American serving with the Royal Canadian Air Force, composed "High Flight." He was born in Shanghai, China in 1922, the son of missionary parents, Reverend and Mrs. John Gillespie Magee; his father was an American and his mother was originally a British citizen.

He came to the U.S. in 1939 and earned a scholarship to Yale, but in September 1940 he enlisted in the RCAF and graduated as a pilot. He was sent to England for combat duty in July 1941.

In August or September 1941, Pilot Officer Magee composed "High Flight" and sent a copy to his parents. Several months later, on December 11, 1941 his "Spitfire" airplane collided with another plane over England and Magee, only 19 years of age, crashed to his death. His remains are buried in the churchyard cemetery at Scopwick, Lincolnshire.

This can be found on the website:

<http://www.wpafb.af.mil/museum/history/prewwii/jgm.htm>

PREFACE

Nicolaus Copernicus and Galileo Galilei developed the scientific knowledge that became the underpinning of spaceflight. Edward Everett Hale in “The Brick Moon” and Jules Verne in “From the Earth to the Moon” dreamed and wrote about it. But finally in the last half of the twentieth century, it was the Americans and the Soviet Russians, locked in the throes of the Cold War, who accomplished it. A good case can be made that when historians look back at the twentieth century, the initial efforts of humankind to slip “the surly bonds of Earth” will play a dominant role.

Today, we call the sixteenth century the “Age of Exploration” because by combining the fore-and-aft sail rig of Arab dhows with the sturdy hull of the Baltic cog, the “caravel” was created that could safely sail all the oceans of the world. Thus, in the final years of the fifteenth century, Bartolomeo Diaz, Christopher Columbus, and Vasco da Gama opened astonishing new vistas using the caravels. In less than a century after their epochal voyages, the geography of the Earth was essentially understood and things were forever changed.

Today, because of the advent of rocket technology, we stand at the threshold of sending humans to Mars as well as to other places in the Solar System. We are within a decade of sending people back to our own Moon to establish permanent stations to exploit lunar resources and to create staging bases for the large-scale exploration of the Solar System. As was the case half a millennium ago, things will change forever when this is done.

We have both been involved in this initial exploratory effort in an intimate way. One of us (Richard H. Truly) has actually flown in space and both of us have participated in and led the organizations established in the United States to conduct space exploration. Both of us have also been touched by the brutal wars of the twentieth century, and we therefore know how these have influenced the lives of people all over the world as well.

The idea of this *Encyclopedia of Space Science and Technology* was conceived late in 1997 when one of us (Hans Mark) had a conversation with Dr. Edmund H. Immergut, who has had a long and distinguished career in scientific publishing and in the production of encyclopedias. He believed that the enterprise of space exploration was far enough along – 40 years after the first orbital flight of Sputnik I – that a good technical encyclopedia on the subject would be timely and appropriate. In developing the ideas for the encyclopedia, the following principles were established.

- The encyclopedia would be written at a high technical level, i.e., for an audience of technically literate people who were not experts in space science or technology.
- The encyclopedia would contain articles that would describe the technology of space exploration as well as the scientific results and their applications.
- The authors who would be selected to write articles would be people who are, or have been, active participants in enterprise of space exploration.

- The encyclopedia would be international and would attempt to capture the spirit that animated the enterprise for the past half century.
- The encyclopedia would have a broadly based editorial board whose members would help to select authors and assist in passing judgment on the quality of the work.

It is our hope that we have largely adhered to these principles. The *Encyclopedia of Space Science and Technology* consists of nearly 80 articles organized under eight separate categories. There is an appropriate index and a table of contents that should make it easy for readers to find the topic of interest for which they are searching.

Throughout this work, both of us have enjoyed working with old and new colleagues. We would like to extend our appreciation to everyone who participated in this effort, first and foremost, our authors for their contributions, our Associate Editors, Drs. Milton A. Silveira and Michael I. Yarymovych, and all the members of our Editorial Board for their participation and advice. Finally, special thanks are due to our Managing Editor, Ms. Maureen A. Salkin, for her tireless and highly diplomatic efforts to keep things rolling so that we can now all see the final result.

Richard H. Truly
Golden, Colorado

Hans Mark
Austin, Texas

CONTRIBUTORS

Brian Allen, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets

R.C. Anderson, *California Institute of Technology, Pasadena, California*, Pathfinder Mission to Mars

Kenneth M. Baldwin, *University of California, Irvine, California*, Muscle Loss in Space: Physiological Consequences

Vyacheslav M. Balebanov, *Russian Academy of Sciences, Institute of Space Research, Russia*, Plasma Thrusters

V.A. Bartenev, *Scientific-Production Association of Applied Mechanics, Russia*, Communication Satellite Development in Russia

Alexander T. Basilevsky, *Vernadsky Institute of Geochemistry and Analytical Chemistry, Russian Academy of Sciences, Moscow, Russia*, Exploration of the Moon by Soviet Spacecraft; Venus Missions

J.R. Beattie, *Westlake Village, California*, Rockets, Ion Propulsion

Robert R. Bennet, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets

George A. Berkley, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets

Jon H. Brown, *Fort Worth, Texas*, Spacecraft Guidance, Navigation and Control Systems

Boris Chertok, *ENERGIA Space Association, Russia*, Sputnik 1: The First Artificial Earth Satellite

Edward L. Chupp, *University of New Hampshire, Durham, New Hampshire*, Sun

Anita L. Cochran, *The University of Texas McDonald Observatory, Austin, Texas*, Comets

Aaron Cohen, *NASA – Lyndon B. Johnson Space Center, Space Shuttle Orbiter Project Office, Houston, Texas*, Space Shuttle Orbiter

Richard J. Cohen, *Harvard University—Massachusetts Institute of Technology, Cambridge, Massachusetts*, Cardiovascular System in Space

Glenn D. Considine, *Westfield, Massachusetts*, Mars

Douglass B. Cook, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets

Robert L. Crippen, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets

F.A. Cucinotta, *NASA Johnson Space Center, Houston, Texas*, Space Radiation

Alexander F. Dedus, *Russian Aviation and Space Agency, Russia*, Russian Spaceports

J.F. Dicello, *Johns Hopkins University School of Medicine, Baltimore, Maryland*, Space Radiation

Steven D. Dorfman, *Hughes Electronics Corporation, Los Angeles, California*, Commercial Applications of Communications Satellite Technology; Communications Satellites, Technology of

Timothy E. Dowling, *University of Louisville, Louisville, Kentucky*, Jupiter

V. Reggie Edgerton, *University of California, Los Angeles, California*, Muscle Loss in Space: Physiological Consequences

Alexander N. Egorov, *Yu.A. Gagarin Cosmonaut Training Center, Russia*, Cosmonauts Selection and Preparation

Gabriel Elkaim, *Stanford University, Stanford, California*, Global Positioning System (GPS)

Maxime A. Faget, *NASA-Johnson Space Center, Houston, Texas*, U.S. Manned Spaceflight: Mercury to the Shuttle

Dale Fenn, *Orbital Sciences Corporation, Dulles, Virginia*, Air and Ship-Based Space Launch Vehicles

Harold B. Finger, *National Aeronautics and Space Administration and Atomic Energy Commission, Washington, D.C.*, Nuclear Rockets and Ramjets

Uwe Fink, *Lunar and Planetary Lab University of Arizona, Tucson, Arizona*, Saturn System

Charles T. Force, *Tracy's Landing, Maryland*, Earth-Orbiting Satellites, Data Receiving and Handling Facilities

Marvin Glickstein, *Pratt & Whitney, Palm Beach, Florida*, Liquid-Fueled Rockets

Teresa Gomez, *NASA Johnson Space Center, Houston, Texas*, Astronauts and the People who Selected Them: A Compendium

L. Gorshkov, *ENERGIA RSC, Russia*, Russian Space Stations

- Robert P. Graham**, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets
- Anatoly I. Grigoriev**, *Institute of Biomedical Problems, Russian Academy of Sciences, Moscow, Russia*, Biomedical Support of Piloted Spaceflight; Space Life Sciences
- Herbert Gursky**, *Naval Research Laboratory, Washington, DC*, Science from Sounding Rockets
- Martin Harwit**, *Cornell University, Ithaca, New York*, Astronomy–Infrared
- W. Michael Hawes**, *NASA, Washington, District of Columbia*, International Space Station
- Clark W. Hawk**, *Madison, Alabama*, Rocket Propulsion Theory
- Steven A. Hawley**, *NASA Johnson Space Center, Houston, Texas*, Human Operations in Space During the Space Shuttle Era
- Hans E.W. Hoffmann**, *ORBCOMM LLC, Dulles, Virginia*, Spacelab
- Stephen Horan**, *New Mexico State University, Las Cruces, New Mexico*, Earth-Orbiting Satellites, Data Receiving and Handling Facilities
- Ross M. Jones**, *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California*, Planetary Exploration Spacecraft Design
- Russell Joyner**, *Pratt & Whitney, Palm Beach, Florida*, Liquid-Fueled Rockets
- Joseph Kerwin**, *Houston, Texas*, Skylab
- Joseph J. Klinger**, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets
- Petr I. Klimuk**, *Yu.A. Gagarin Cosmonaut Training Center, Russia*, First Flight of Man in Space
- Stanislav Nikolaevich Konyukhov**, *M.K. Yangel' Yuzhnoye State Design Office, Dniepropetrovsk, Ukraine*, Conversion of Missiles into Space Launch Vehicles
- Jean Kovalevsky**, *Cerga-Observatoire de la Côte d'Azur, Grasse, France*, Optical Astrometry from Space
- A.G. Kozlov**, *Scientific-Production Association of Applied Mechanics, Russia*, Communication Satellite Development in Russia
- Alexander N. Kuznetsov**, *Russian Aviation and Space Agency, Russia*, Russia's Launch Vehicles; Russian Spaceports
- James W. Layland**, *California Institute of Technology, Pasadena, California*, Deep Space Network, Evolution of Technology
- David S. Leckrone**, *NASA, Goddard Space Flight Center, Greenbelt, Maryland*, Hubble Space Telescope
- James R. Lesh**, *California Institute of Technology, Pasadena, California*, Deep Space Network, Evolution of Technology
- John S. Lewis**, *University of Arizona, Tucson, Arizona*, Space Resources, Occurrence and Uses
- Wah L. Lim**, *Hughes Electronics Corporation, Los Angeles, California*, Commercial Applications of Communications Satellite Technology
- Glynn S. Lunney**, *Houston, Texas*, NASA Mission Operation Control Center at Johnson Space Center
- Ronald W. Lyman**, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets
- Dmitry K. Malashenkov**, *Institute of Biomedical Problems, Russian Academy of Sciences, Moscow, Russia*, Biomedical Support of Piloted Spaceflight; Space Life Sciences
- Jerry W. Manweiler**, *Fundamental Technologies LLC, Lawrence, Kansas*, Interplanetary Medium
- Hans Mark**, *Austin, Texas*, Evolution of U.S. Expendable Launch Vehicles
- Ian R. McNab**, *The University of Texas at Austin, The Institute for Advanced Technology, Austin, Texas*, Electromagnetic Propulsion
- Valeriy A. Menshikov**, *Khrunichev Space Center, Moscow, Russia*, Global Navigation Satellite System; Military Use of Space
- Jerome H. Molitor**, *Westlake Village, California*, Rockets, Ion Propulsion
- Vasily I. Moroz**, *Space Research Institute, Russian Academy of Sciences, Moscow, Russia*, Exploration of Mars by the USSR; Venus Missions
- Alexey I. Morozov**, *Russian Science Center, Kurchatov Institute, Russia*, Plasma Thrusters
- David Morrison**, *NASA Ames Research Center, Moffett Field, California*, Asteroids; Astrobiology

- Adam L. Mortensen**, *USAF, USSPACE/SIOE-r, Colorado Springs, Colorado*, Military Ground Control Centers, United States
- Douglas J. Mudgway**, *California Institute of Technology, Pasadena, California*, Deep Space Network, Evolution of Technology
- F. Robert Naka**, *CERA, Incorporated, Concord, Massachusetts*, Space Programs Related to National Security
- John J. Neilon**, *Cocoa Beach, Florida*, Eastern Launch Facilities, Kennedy Space Center
- Robert M. Nelson**, *Jet Propulsion Laboratory, Pasadena, California*, Mercury
- R. Steven Nerem**, *University of Colorado, Colorado Center for Astrodynamics Research, Boulder, Colorado*, Earth Orbiting Satellite Theory
- Arleigh P. Neunzert**, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets
- Arnauld Nicogossian**, *NASA Headquarters, Washington, D.C.*, Biological Responses and Adaptation to Spaceflight: Living in Space—an International Enterprise
- Bradford Parkinson**, *Stanford University, Stanford, California*, Global Positioning System (GPS)
- Billy H. Prescott**, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets
- C. Paul Pulver**, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets
- Craig D. Ramsdell**, *Beaumont Hospital, Royal Oak, Michigan*, Cardiovascular System in Space
- P. Krishna Rao**, *National Oceanic and Atmospheric Administration, Silver Spring, Maryland*, Weather Satellites
- Lawrence L. Rauch**, *California Institute of Technology, Pasadena, California*, Deep Space Network, Evolution of Technology
- John C. Ries**, *The University of Texas at Austin, Center for Space Research, Austin, Texas*, Precision Orbit Determination for Earth Observation Systems
- Robert Rosen**, *NASA Ames Research Center, Moffett Field, California*, Liquid-Fueled Rockets
- Duane L. Ross**, *NASA Johnson Space Center, Houston, Texas*, Astronauts and the People who Selected Them: A Compendium
- Roland R. Roy**, *Brain Research Institute, University of California, Los Angeles, California*, Muscle Loss in Space: Physiological Consequences
- Roald Sagdeev**, *University of Maryland, College Park, Maryland*, Vega Project
- Donald R. Sauvageau**, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets
- H.H. Schmitt**, *University of Wisconsin—Madison, Wisconsin*, Apollo 17 and the Moon
- B.E. Schutz**, *University of Texas at Austin, Center for Space Research, Austin, Texas*, Size and Shape of Earth from Satellites
- Yuri P. Semyonov**, *ENERGIA RSC, Russia*, Russian Space Stations
- William T. Shearer**, *Texas Children's Hospital, Houston, Texas*, Immunology and Infection in Space
- Milton A. Silveira**, *NASA Johnson Space Center, Houston, Texas*, Space Shuttle Orbiter; U.S. Manned Space Flight: Mercury to the Shuttle
- S. Fred Singer**, *The Science & Environmental Policy Project (SEPP), Arlington, Virginia*, Weather Satellites
- G.M. Solovyev**, *Khrunichev Space Center, Russia*, Global Navigation Satellite System
- Gerald Sonnenfeld**, *Morehouse School of Medicine, Atlanta, Georgia*, Immunology and Infection in Space
- Yu.B. Sosyurka**, *Yu.A. Gagarin Cosmonaut Training Center, Russia*, Cosmonauts Selection and Preparation
- James Spilker**, *Stanford University, Stanford, California*, Global Positioning System (GPS)
- Paul D. Spudis**, *Lunar and Planetary Institute, Houston, Texas*, Moon
- Lawrence A. Sromovsky**, *University of Wisconsin, Madison, Wisconsin*, Uranus and Neptune
- William Stoney**, *Mitretek Corporation, Reston, Virginia*, Civil Land Observation Satellites
- Byron D. Tapley**, *The University of Texas at Austin, Center for Space Research, Austin, Texas*, Precision Orbit Determination for Earth Observation Systems
- Jill Tarter**, *SETI Institute, Mountain View, California*, Extraterrestrial Life, Searching for

Roger Vignelles, *Corbeil-Essonnes, France*, Ariane Rocket Program

G.I. Vorobyov, *Yu.A. Gagarin Cosmonaut Training Center, Russia*, First Flight of Man in Space

Steven R. Wassom, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets

Martin C. Weisskopf, *NASA – Marshall Spaceflight Center, Huntsville, Alabama*, Chandra X-ray Observatory

Michael Werner, *Jet Propulsion Laboratory, Pasadena, California*, Astronomy-Infrared

Nicholas J. Whitehead, *Thiokol Propulsion, Inc., Brigham City, Utah*, Solid Fuel Rockets

Simon P. Worden, *USAF, USSPACE/SIOE-r, Colorado Springs, Colorado*, Military Ground Control Centers, United States

V.I. Yaropolov, *Yu.A. Gagarin Cosmonaut Training Center, Russia*, Cosmonauts Selection and Preparation

Michael I. Yarymovych, *Boeing Space and Communications (Retired), Seal Beach, California*, Evolution of U.S. Expendable Launch Vehicles

Laurence R. Young, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Artificial Gravity

Eliot Young, *Southwest Research Institute, Boulder, Colorado*, Pluto and Charon

Leslie Young, *Southwest Research Institute, Boulder, Colorado*, Pluto and Charon

Alexander V. Zakharov, *Space Research Institute, Russian Academy of Sciences, Moscow, Russia*, Exploration of Mars by the USSR

ENCYCLOPEDIA OF

SPACE SCIENCE
AND
TECHNOLOGY

VOLUME 1

A

AIR AND SHIP-BASED SPACE LAUNCH VEHICLES

Introduction

In 1957, the Soviet Union placed the first man-made object in orbit around the earth. Since then, numerous launch vehicles have been developed to improve the performance, reliability, and cost of placing objects in orbit. By one estimate, roughly 75 active space launch vehicles either have established flight records or are planning an inaugural launch within the year. This does not include the numerous launch vehicles from around the world that are no longer operational such as the Jupiter, Redstone, Juno, Saturn, Scout, Thor, Vanguard, and Conestoga family of rockets from the United States or the N-1 from the former Soviet Union, to name just a few. Despite the many differences among all of these launch vehicles from both past and present, one common element can be found in all but four of them: they are ground-launched. Of the four exceptions, two are air-launched (NOTSNIK and Pegasus), one is ship-launched (Sea Launch), and one is submarine-launched (Shtil). It is important to keep in mind that numerous air-launched and ship-launched suborbital launch systems are in use by militaries, commercial entities, and educational institutions. However, the four mentioned are the only mobile launch systems that can place objects into a sustainable Earth orbit.

Mobile Space-Launched Vehicles

Project Pilot (NOTSNIK). NOTSNIK is the oldest and, until recently, the least well known of the four mobile space-launched systems. Following the launch of Sputnik by the Soviet Union, President Eisenhower's administration elicited proposals to launch a satellite into orbit. The Naval Ordinance Test

Station (NOTS) located at China Lake in California proposed launching a rocket from a jet fighter (1). The idea is the same as that of the current Pegasus vehicle: reduce the amount of energy needed to place a payload into orbit by launching it above the denser portion of the atmosphere. In this fashion, the engineers at NOTS designed a vehicle from existing rocket motors that could place a 2-pound satellite in a 1500-mile-high orbit. The engineers recognized the energy savings from such a launch concept and also the utility of such a flexible platform. Launching from a jet fighter could, theoretically, place a satellite into any orbit from anywhere in the world at any time.

The U.S. Navy accepted the proposal from NOTS in 1958, by some accounts as a safety net in the event that the ongoing Vanguard project was unsuccessful. The program was officially called Project Pilot, but the engineers at NOTS preferred the name NOTSNIK in direct reference to the Soviet satellite that was currently orbiting above them and the rest of the world. A Douglas Aircraft F4D-1 Skyray was the carrier aircraft for the rocket and consequently was considered the first stage. The second and third stages were modified antisubmarine missiles. The final stage was taken from a Vanguard rocket. The entire launch vehicle measured a mere 14 feet in length and had four fins at the aft end that provided a span of 5 feet.

The NOTSNIK was launched six times from an altitude of about 41,000 ft. Four of those launches ended in known failures. However, the results of two have never been verified. Some in the program insist that they achieved their goal of placing the small payload of diagnostic instruments in orbit. At least one ground station in New Zealand picked up a signal in the right place at the right time. However, confirmation that the signal was from the NOTSNIK payload was never established. Even the possibility of a success was veiled in secrecy for more than 40 years for, by all accounts, two critical reasons. The first was that in the days following the early embarrassments of Vanguard, the Eisenhower administration did not want to claim success unless it was absolutely certain. The second reason was that a mobile air-launched system that could reach orbit had extremely appealing military applications. However, the tactical advantages of such a system were far outweighed by the strategic consequences, as stated in the Antiballistic Missile (ABM) Treaty between the United States and the former Soviet Union that was concluded in 1972 (2):

Further, to decrease the pressures of technological change and its unsettling impact on the strategic balance, both sides agree to prohibit development, testing, or deployment of sea-based, air-based, or space-based ABM systems and their components, along with mobile land-based ABM systems. Should future technology bring forth new ABM systems 'based on other physical principles' than those employed in current systems, it was agreed that limiting such systems would be discussed, in accordance with the Treaty's provisions for consultation and amendment.

Pegasus. Roughly 30 years later, while NOTSNIK remained an official government secret, the idea of launching payloads into space from an airborne platform was revisited in the form of the Pegasus launch vehicle. The driving forces behind NOTSNIK and Pegasus were essentially the same. An air-launched space vehicle provides several advantages compared with ground-based counterparts.

As an example, Pegasus is launched at an altitude of 39,000 ft, which is above a significant portion of the atmosphere. As mentioned, with NOTSNIK, this eliminates the need for extra performance that would otherwise be needed to overcome atmospheric forces. This also implies that the structural components of the vehicle can be lighter, which improves the efficiency of the rocket as a whole. The energy required from the launch vehicle is also reduced by the speed already achieved by the carrier aircraft. An air-launched system also allows applying more of the impulse of the first stage along the velocity vector. This is a more efficient use of the vehicle's energy than that of ground-launched vehicles that must first apply the thrust almost perpendicular to the velocity vector already imparted by Earth's rotation. These factors combine to produce a requirement for a velocity increment that is on the order of 10% less than a comparable ground-launched rocket.

The Pegasus vehicle is a winged, three-stage, solid rocket booster (Fig. 1). It is the first space-launched vehicle developed solely with commercial funding. Three versions have been developed and flown over the years: Standard, Hybrid, and XL. The XL is the only vehicle within the Pegasus family currently in production. The XL is roughly 10,000 lbm heavier than the Standard or Hybrid models and is roughly 6 ft longer. Because the XL extends farther aft beneath the L-1011 carrier aircraft, the port and starboard fins become an obstacle to the landing gear doors. To correct this problem, the port and starboard fins were

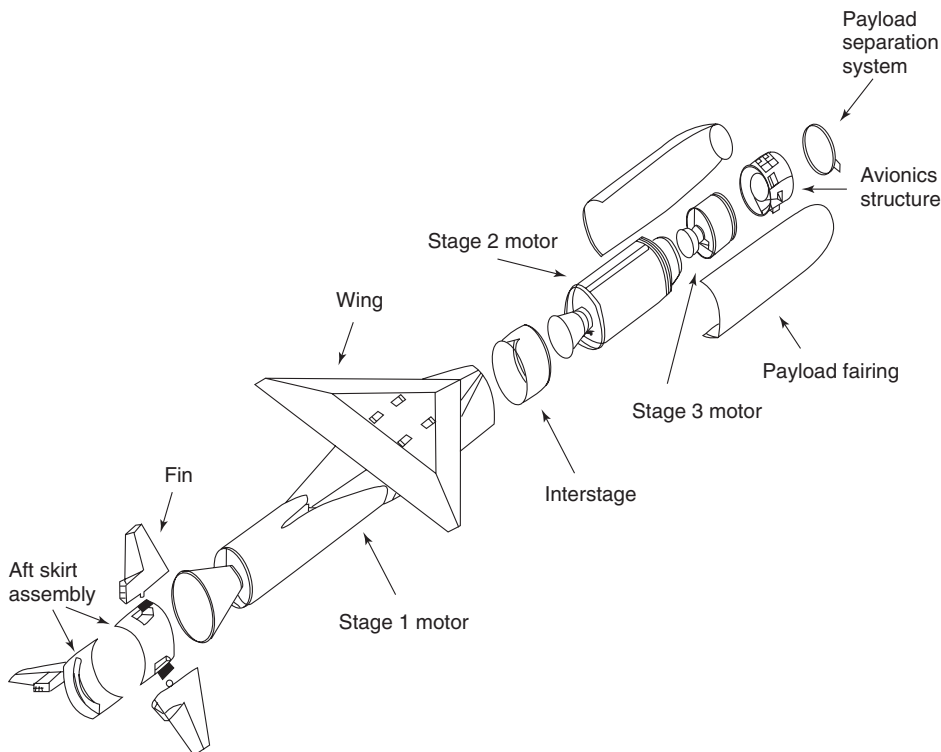


Figure 1. Disassembled version of standard Pegasus launch vehicle.

modified to include an anhedral of 23° . To maintain commonality between the various members of the Pegasus family of vehicles, the same anhedral was introduced into the Standard vehicle, which was then given the designation Pegasus Hybrid. Other than the anhedral of the fins, the Standard and Hybrid vehicles are exactly the same. The Standard, the first Pegasus vehicle built, was flown on six missions. The Hybrid vehicle has flown four times. The XL vehicle has flown 21 times. Of 31 Pegasus launches, only three missions failed to reach orbit.

The Pegasus XL was designed and developed to provide increased performance above and beyond that provided by the Standard and Hybrid vehicles. A typical Pegasus XL vehicle weighs roughly 51,000 lbm at launch, is 55.4 ft long and 50 inches in diameter, and the wingspan is 22 ft (3). At launch, the Pegasus XL is carried aloft by the company's carrier aircraft, a modified L-1011, which originally saw commercial service with Air Canada. The vehicle is dropped from an altitude of 39,000 ft at Mach 0.8. Five seconds after release from the L-1011, the first stage ignites and the vehicle's on-board flight computer continues the sequence of events that eventually lead to orbital insertion. The brief coast period between drop and stage one ignition is designed to provide a safe distance between the L-1011 and the launch vehicle.

The Pegasus Standard vehicle was originally dropped from a NASA-owned and operated B-52. The Pegasus vehicle was attached to one of the pylons underneath the starboard wing much in the same manner as the early supersonic and hypersonic test vehicles such as the X-15. For a variety of reasons, Orbital purchased and modified the L-1011 to facilitate all future launches.

Unlike the B-52 that supported initial Pegasus launches, the L-1011 carries the Pegasus vehicle underneath the fuselage rather than underneath the wing. Once Pegasus is ready to be mated to the carrier aircraft, it is towed from Orbital's integration facility at VAFB to the plane on the Assembly and Integration Trailer (AIT). Regardless of where the launch is to take place, the Pegasus is always integrated and mated to the L-1011 at VAFB. From there, the launch system can travel to any location in the world for launch. There is enough ground clearance for the L-1011 to take off and land with Pegasus attached underneath. However, the added height of the AIT underneath Pegasus requires raising the L-1011 off the ground slightly by hydraulic jacks to mate Pegasus to the carrier aircraft (Fig. 2). While mated to the L-1011, the vertical rudder actually protrudes into the plane's fuselage in a compartment specifically designed for this purpose. When mating the Pegasus to the L-1011, the rudder is usually detached from the Pegasus vehicle and placed inside the housing first. Then the Pegasus is rolled underneath the L-1011 and attached to the rudder and then to the plane. Removing the rudder first minimizes the height to which the L-1011 needs to be raised for the mating process. The entire mating process from rollout to mating takes about 6 hours. Pegasus is attached to the L-1011 using four hooks on the center box of the wing and a fifth hook on the forward portion of the vehicle. The inside of the airplane has been stripped of all unnecessary equipment and hardware. Up front in what would normally be the first class cabin are eight seats for personnel during ferry flights from VAFB to the launch site of interest and two computer stations from which personnel can monitor the health of the vehicle and the payload. The rest of the interior of the cabin has been completely gutted. Access to the rear portion of the aircraft cabin is obtained through a galley door.



Figure 2. Fully assembled Pegasus launch vehicle being mated to the L-1011 aircraft. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Unlike most other launch vehicles in the U.S. fleet, the Pegasus launch vehicle is integrated horizontally on the AIT (Fig. 3). Horizontal integration facilitates easy access to the vehicle and eliminates the need for high bays and large cranes. Components are received as needed either from groups within Orbital Sciences or from outside vendors. To ensure that all of the major flight hardware and software is thoroughly tested before flight, Pegasus, like many other vehicles, is subjected to a series of “fly to orbit” simulations at various stages of the integration process. Four flight tests are normally performed. The first tests the three stages individually. The second test is conducted after the three stages are electrically mated together. The third test is performed after the three stages are electrically and mechanically mated and the stack is electrically mated to the payload. The fourth and final flight test is performed once the payload has been mechanically mated to the rest of the vehicle and the half of the fairing that includes the pyro devices necessary for jettisoning the shroud is



Figure 3. Horizontal integration of Pegasus launch vehicle. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

electrically mated. These tests are intended to verify that various systems function and also respond as expected to known disturbances. If the inertial measurement unit (IMU) onboard receives data to indicate that an unexpected attitude change has occurred, will the fins or thrust vector control systems respond accordingly? Are all the commands to the various subsystems appropriate, and do those subsystems respond appropriately? Once the Pegasus vehicle has been mated to the L-1011 carrier aircraft, one last test is performed, called the Combined Systems Test (CST). This test verifies that the launch vehicle and the carrier aircraft are communicating as expected. This is particularly important since the vehicle's health can be monitored both from telemetry that is broadcast from the vehicle to the ground via antennas on Pegasus and also by the computer stations inside the L-1011 via hardwired electrical connections. More importantly, some data and commands are sent to the Pegasus vehicle before launch. The only method currently available for accomplishing this transfer of data is through the electrical connections between the Pegasus vehicle and the carrier aircraft.

To be fully mobile, the Pegasus launch system must also be fully self-contained. Except for those services provided by the range (such as radar coverage), the L-1011 can transport all of the equipment required to support a launch of Pegasus, including, of course, Pegasus itself (Fig. 4). Some launches take place off the coast of California where the Western Range (based at VAFB) is the lead range. In these instances, no ferry flight is required. The L-1011 simply takes off from VAFB and flies to the designated drop point roughly 100 nmi out to sea. The checklist that is processed in the control room on the day of launch requires about 4 to 5 hours to complete. The L-1011 usually takes off an hour before the scheduled launch time. If all systems are "go," as determined by the mission team members in the control room, the launch conductor on the ground commands the pilot of the L-1011 to drop the Pegasus from the carrier aircraft.

Shtil. In a classic example of turning swords into plowshares, the Russian Navy developed a satellite delivery system for nonmilitary applications that uses a submarine-launched. The SS-N-23 (NATO's designation) is a three-stage



Figure 4. L-1011 aircraft taking off with Pegasus. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

liquid-fueled vehicle that can deliver small satellites to low Earth orbit. Very little is known about this launch vehicle service including performance to various altitudes and inclinations. What is known is that two satellites belonging to the Technical University of Berlin were successfully launched in 1998 from a Russian submarine for the stunningly low price of \$150,000 (4). Some sources indicate that the typical commercial price for a Shtil launch is actually in the neighborhood of \$500,000. There are two possible reasons for the low cost of a Shtil launch. The first is that more than 200 missiles have already been produced by the Russian military. There is also speculation that offering commercial launch services provides a way to maintain proficiency in launching missiles without using precious military funding. One disadvantage of this system is that the Shtil vehicle likely does not have enough performance to achieve circular orbits in the medium to high Low Earth Orbit (LEO) altitudes (4). This is a direct result of the Shtil's heritage as a ballistic missile first and foremost.

Sea Launch. The most recent mobile launch system is the Sea Launch vehicle which is launched from a converted oil-drilling platform along the equator (Fig. 5). Sea Launch is both the name of the launch vehicle and the name of the international joint venture that provides the launch services. The partnership is comprised of Boeing, KB Yuzhnoye of Ukraine, which provides the two Zenit stages, and RSC Energia of Russia, which provides the Block DM-SL upper stage. The launch vehicle and payload integration takes place at the vehicle's home port of Long Beach, California. Once integration is complete, the launch vehicle is loaded onto the converted oil-drilling platform and towed to a predetermined

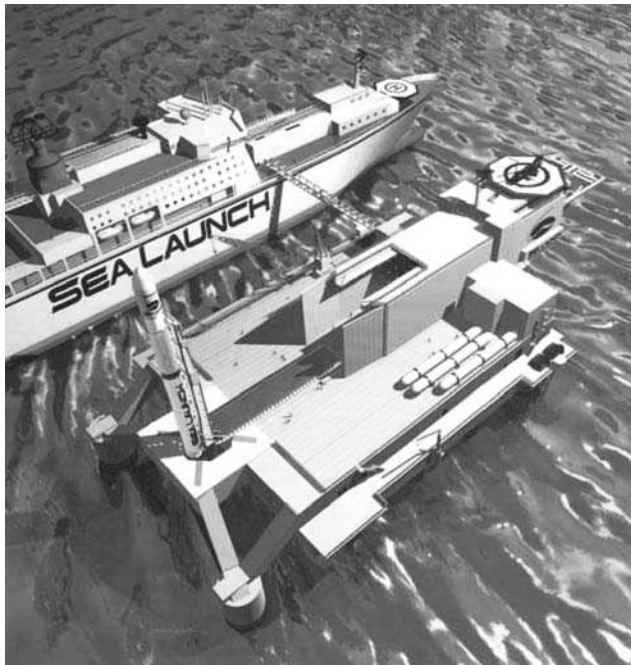


Figure 5. Computer simulation of Sea Launch. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

launch location at the equator, specifically 154° West. Once on site, the Zenit 3SL is raised into its launch attitude (vertical) and launched. A second ship that houses mission personnel and the control room monitors the launch from nearby. The vehicle itself is a little less than 200 ft long and roughly 13 ft in diameter. The performance to Geosynchronous Transfer Orbit (GTO) is approximately 5250 kg (4). “In terms of spacecraft mass in final orbit, this would be equivalent to approximately 6000 kg of payload capability if launched from Cape Canaveral, because the spacecraft does not need to perform a plane change maneuver during the Geosynchronous Earth Orbit (GEO) circularization burn” (5).

There are three key phases in the integration of a Sea Launch vehicle (5). Phase I takes place in the Payload Processing Facility (PPF). This phase includes receipt of the spacecraft, processing of the spacecraft, testing, and enclosure within the payload fairing. Phase II takes place on the Assembly and Command Ship (ACS). This entails mating the encapsulated spacecraft to the launch vehicle and testing the integrated stack. Phase III takes place on the Launch Platform (LP) once the vehicle has been transferred from the ACS. While still in port, the integrated launch vehicle is raised to its vertical launch attitude so that a series of tests can be conducted. The launch vehicle is then lowered back into a horizontal position, stored in an environmentally controlled room, and transported to the equator while on board the launch platform. At the launch site, the launch vehicle is rolled out to the launch pad, raised to a vertical attitude again, and fueled. The launch is performed by an automated system and monitored by the Assembly and Command Ship which is moved for launch to a distance 6.5 km away (Fig. 6).

The Assembly and Command Ship for Sea Launch serves as the launch vehicle integration and testing facility. In addition to acting as the temporary home for launch crews, the ship also houses the Launch Control Center (LCC) and the equipment necessary to track the initial ascent of the rocket. Unlike the



Figure 6. Sea Launch successfully lifts DIRECTV 1-R satellite into orbit. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Pegasus carrier aircraft that was modified after serving in a different capacity, the ACS was designed and constructed specifically to suit the unique requirements of Sea Launch. The ship is roughly 660 ft long and 110 ft in beam and has an overall displacement of approximately 30,830 tonnes.

The rocket assembly facility is on the main deck of the ACS where the launch vehicle integration takes place. This activity is conducted before setting sail for the equator and simultaneously with spacecraft processing. After the spacecraft has been satisfactorily processed, it is encapsulated and transferred to the rocket assembly compartment, where it is mated to the launch vehicle. Following integration and preliminary testing, the integrated launch vehicle is transferred to the launch platform. Then both ships begin the journey to the equator, which takes roughly 12 days.

The launch platform has all of the necessary systems for positioning and fueling the launch vehicle, as well as conducting the launch operations. Once the launch vehicle has been erected and all tests are complete, personnel are evacuated from the launch platform to the ACS using a link bridge between the vessels or a helicopter. Redundant radio-frequency links between the vessels permit personnel on the ACS to control all aspects of the launch, even when the command ship has retreated to a safe distance before launch. The launch platform, which was converted from an oil drilling platform, is very stable. It is supported by a pair of large pontoons and is propelled by a four-screw propulsion system (two in each aft lower hull). Once at the launch location, the pontoons are submerged to a depth of 70.5 ft to achieve a more stable attitude for launch, level to within approximately 1° .

Advantages of Mobile Space-Launched Systems

NOTSNIK, Pegasus, Sea Launch, and Shtil were never intended to replace the existing fleet of ground-launched rockets. Rather, they effectively supplement the existing worldwide capability by providing additional services to a targeted market of payloads that benefit greatly from the mobility and flexibility of these unique space-launch systems. These vehicles can provide services similar to ground-launched vehicles for payloads within their weightclass. In fact, all four vehicles have fixed launch locations for standard services. For example, Pegasus uses the launch location of 36°N , 237°E for all high-inclination missions that originate from VAFB. In this regard, the mobile launch systems are no different from ground-launched vehicles in that they repeatedly launch from a fixed location, albeit a location that is not on land. However, they can also offer services and performance that avoid many of the restrictions inherent in being constrained to a particular launch site. Few of those restrictions are trivial. They include inclination restrictions, large plane changes required to achieve low-inclination orbits from high-latitude launch sites, large plane changes required to transfer from GTO to GEO when launching from certain ranges, and low-frequency launch opportunities for missions that require phasing such as those involving a rendezvous with another spacecraft already in orbit.

Inclination Restrictions. Inclination restrictions stem from range safety considerations. To understand these restrictions fully, it is first necessary to

understand two concepts: (1) transfer orbits and (2) instantaneous impact-point tracks.

Transfer Orbits. Transfer orbits are intermediate orbits established by the various stages of a launch vehicle that provide a path to the final desired orbit. The transfer orbits for early stages are mostly suborbital, meaning that some portion of the orbit intersects Earth's surface. The most efficient way to transfer between two orbits is to apply thrust at opposite apses. An application of thrust in the right direction at the perigee of the initial orbit will raise the apogee. Coasting to the new apogee and applying thrust (again in the appropriate direction) at this apsis will then raise the perigee. This provides a stair-step approach to raising the altitude of a vehicle's orbit. The ascent of a launch vehicle from launch to orbit follows a similar trend with one critical caveat. The impulse of initial stages is usually not sufficient, individually, to raise the perigee above Earth's surface. This means that using the optimal Hohmann transfer approach would bring the launch vehicle back to Earth before another transfer burn could be made. As a result, initial launch vehicle stages usually apply their thrust at places within a transfer orbit other than the apses and usually always on the ascending side of the orbit.

Consider a modest three stage, ground-launched rocket launching into a circular low Earth orbit as an example. Before launch, the vehicle is effectively sitting at the apogee of an orbit (Fig. 7). If the surface of Earth were not present to support the rocket, it would be drawn downward along a path that would take it closer and closer to Earth's center before swinging back to an apogee altitude equal to the radius of Earth. This is essentially the first of several transfer orbits and the rocket has not even been launched. When the rocket lifts off, it applies its thrust at an apsis, but in a direction that is perpendicular to the initial velocity vector of the rocket, which itself is in the direction of Earth's rotation. During the first burn, the vehicle slowly tilts over so that the thrust is applied in a direction that is increasingly parallel to Earth (Fig. 8). This has the effect of increasing both the apogee and perigee. The perigee will most likely still be suborbital at the end of the burn. The apogee will be increased sufficiently that the launch vehicle

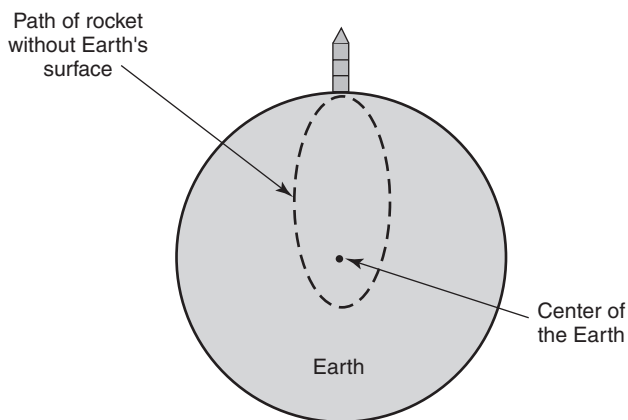


Figure 7. Path of rocket without Earth's surface. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

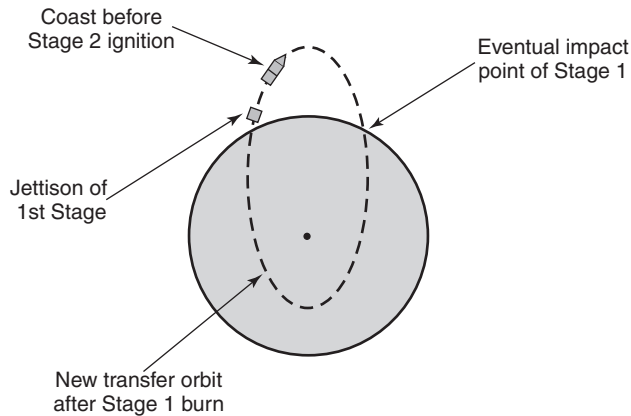


Figure 8. Path of rocket after launch. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

can coast up to a location near the new apogee, following the first stage burnout, and ignite the second stage. The key consideration here is that the second stage will be ignited near but not at the apogee. Again, this is not the most energy-efficient way to transfer orbits, but it is necessary because the opposite apsis is still below Earth's surface, and the second stage may not have sufficient impulse to raise it above the atmosphere. Igniting the second stage at a location other than the apogee again has the effect of raising both the perigee and the apogee. In this case, because only one stage is left, the burn is designed to raise the apogee to the desired altitude of the final orbit. After the second stage burns out, the vehicle coasts up to the new apogee and ignites the third stage. This will raise the perigee up to the final orbit altitude without changing the altitude of the apogee.

Impact-Point Tracks. By always burning on the ascending side of the trajectory and iteratively raising the apogee while the transfer orbit remains suborbital, anything jettisoned before the final burn will reenter the atmosphere and either burn up or impact Earth's surface. As the burn of each stage progresses, the point at which the transfer orbit intersects the Earth extends farther and further downrange until, at some point late in the final burn, there is no longer a point of intersection. These points of intersection comprise the instantaneous impact-point track. Clearly, as the vehicle is coasting, the instantaneous impact point does not change. Conversely, during a motor burn, it is constantly changing and each point represents the location of impact on Earth if, in fact, the thrust were to be instantly terminated either by design or due to some sort of failure. It is this impact-point track and the need for it to avoid populated areas that is a primary source of inclination restrictions from various ranges.

For any rocket launch, whether it be space-based, suborbital, ground-launched, ship-launched, or air-launched, the public-safety considerations that must be satisfied are very stringent. Those stages of a rocket that are jettisoned before reaching orbit should avoid land. And no launch vehicle whose impact-point track nominally crosses land can risk a casualty among the public with a probability of greater than 30 in a million. Calculating the expectation of a

casualty depends on many factors, including the reliability of the launch vehicle (e.g., how many failures it has had in the past), the density of the population being overflown, and the speed with which the instantaneous impact-point track crosses over a populated region. Late in flight, the distance between successive impact points increases dramatically and reduces the risk to the population below. This is why it is generally more permissible to overfly populated regions far downrange than it is early in flight. For instance, the risk to a populated region in Africa from a rocket launched at the Eastern Range would, in general, be less than the risk posed to an area with the same population density overflown in the Caribbean. This is not to say that overflight of any part of Africa is acceptable. There are some extremely high population densities in Africa, especially along the west coast of northern Africa, which are avoided at all costs. And it is this very consideration that constrains the paths of many launch vehicles from the existing ranges.

The key land masses that must be avoided early in flight for vehicles launching from the Eastern Range include the entire eastern seaboard of the United States when launching on an ascending pass (northerly direction) and the Caribbean and South America when launching on a descending pass (southerly direction). For maximum performance from any given launch vehicle, this restricts the range of inclinations achievable from the Eastern Range to between roughly 28.5° and 51° for ascending passes and between 28.5° and 40° for descending passes. Clearly, inclinations outside this range would be achievable if plane changes were instituted, but that has the disadvantage of reducing the maximum available performance for any given launch vehicle. Higher inclinations are available from the Western Range but restrictions still exist there due to Hawaii, islands in the South Pacific, and the western coasts of both North and South America.

When the inclinations from both the Eastern and Western Ranges are combined (assuming direct injection), a block of inclinations is unavailable without plane changes and subsequent reductions in weight-to-orbit capabilities. For small payloads with limited budgets that require an inclination outside what is directly available from the existing ranges, the cost of launching on the heavy-lift launchers that can execute the necessary plane changes can be prohibitive. And reducing launch costs by flying as a secondary or even tertiary payload is advantageous only in the rare event that a primary payload can be found that requires the same final orbit. For these customers, Pegasus and Shtil provide an alternative due to their relatively low cost, mobility, and self-contained launch infrastructure. Sea Launch provides a similar alternative for the heaviest satellites that are intended for either GEO or low Earth orbits.

Plane Changes Required to Achieve Low Inclinations. The inclination of an orbit represents the angle between the equatorial plane and the orbital plane around Earth. This also happens to be similar to the definition of lines of latitude. It is no coincidence then that the maximum latitude of the ground track for any object in space is roughly equivalent to the inclination of the object's orbit. The only reason that the maximum latitude is not exactly equal to the inclination is because Earth is not a perfect sphere. Conversely, this implies that the minimum inclination attainable by a launch vehicle is roughly equivalent to the latitude of the location from which it is launched. The maximum is 180° minus

the latitude of the launch point. This leads to the important conclusion that the only latitude from which all inclinations are directly accessible is 0° (the equator). The Eastern Range is at a latitude of roughly 28.5° . Therefore, the minimum inclination attainable without plane changes is roughly 28.5° . Lower inclinations can be achieved by launching into any available inclination, achieving a preliminary orbit, and then making an inclination correction burn when the satellite is over the equator or at any latitude that is numerically less than the desired inclination. The significant disadvantage of this process is that inclination changes while in orbit require a great deal of energy. The larger the change in inclination required, the more energy must be expended. Depending on the final orbit desired, this usually requires an additional stage to correct the inclination and achieve the final orbit. The most common recipient of this type of orbit maneuver is a satellite headed to geosynchronous orbit. However, there are low Earth orbit payloads that require low inclinations as well. The ability of Pegasus and Shtil to move the drop point to a latitude from which such energy-intensive plane changes would not be required permits smaller launch vehicles to achieve the same orbit from lower latitudes that larger vehicles can achieve from higher latitudes. The difference in cost, complexity, and performance can often mean the difference for some customers between launching or not.

Some launch locations maintained by other countries are at significantly lower latitudes than those in the United States. For some customers, such ranges can provide the necessary services. However, many satellites in the United States, especially government sponsored, are required to contract with a U.S. launch service provider and use a U.S. controlled range.

Phasing. An object's orbit is essentially a locus of points that defines the path of the satellite. Those points define a plane that goes through the center of Earth. To define an object's precise position within an orbit, that plane and every position in it is defined with respect to both Earth and a coordinate system, one of whose axes always points toward the vernal equinox. Every position of a satellite as it orbits Earth is defined in terms of an epoch (time), the semimajor axis, and eccentricity, measured from Earth's center, inclination and argument of perigee, which are both referenced to Earth's equator, and the right ascension of the ascending node, which is referenced to the vernal equinox frame.

A rendezvous between two objects in space involves a series of maneuvers designed to make the orbital elements of both objects the same, hence confirming the fact that they have, in fact, become a single object orbiting Earth. Just as motor burns can raise or lower the perigee or apogee of an orbit or change the inclination, so too can motor burns be used to change every orbital element that defines a satellite's motion. However, changing some of those elements, especially those that require plane changes, requires large amounts of energy, and they are considered "expensive" in the parlance of orbital mechanics. One way to avoid paying the high price of actively changing the orbit of a satellite with a motor burn is to do it passively through the aid of various external forces. Several naturally occurring forces cause every orbital element to change over time. These include atmospheric drag, solar radiative pressure, the gravitational attraction of the Moon, Sun, and planets, and the nonuniform gravitational forces due to Earth's oblateness. These forces can be used to one's advantage when planning a rendezvous mission. However, some changes resulting from these forces can take

a very long time to reach significant levels. This means that the initial differences between the rendezvousing satellite and the target must be initially small to avoid spending too much unproductive time in orbit. This can be accomplished by simply timing the launch appropriately so that at the time of orbital insertion, the satellite that has newly arrived in orbit is very close to the orbital plane of the target satellite.

To accomplish this maneuver, the launch must occur when the target satellite passes almost directly overhead. It also must be passing in the same direction as the intended launch. In other words, if the satellite being launched is to head off in a southerly direction (along the descending pass), the target satellite must be overhead and also on its descending pass as well. Otherwise, the two satellites will end up with right ascensions that are 180° apart which would be excessively expensive (either in terms of time or energy) to correct once in orbit.

For ground-launched vehicles, the wait between successive passes of the target satellite could be as much as several days, depending on the target orbit because the distance between ground tracks on successive passes depends on the period of the orbit, which depends on the orbit's semimajor axis. Clearly, the ground track of an object that requires only 90 minutes to orbit Earth will be more closely spaced than the ground track of an object whose period is several hours. These ground tracks will pass to the east and west of the given launch site on a daily basis, but the distance between the ground track and the launch site will only be minimized by a periodicity of the order of days.

Mobile assets, however, can eliminate the wait by essentially choosing a launch point that is ideally suited for a rendezvous. Instead of waiting for the ground track to come to the launch point, the launch point is moved to the ground track. In this way, the launch opportunities can be reduced from one every two to three days to at least once a day if not twice a day, if the launch vehicles have the flexibility to launch on both ascending and descending passes.

Consider an example of a satellite being launched by a Pegasus XL to rendezvous with a satellite currently in orbit at an altitude of 400 km circular. A normal ground-launched vehicle would require a wait of about 2 days between successive launch attempts. However, the mobility of Pegasus permits two launch opportunities every day, which is graphically represented in Figs. 9 and 10. Two key assumptions need to be kept in mind when viewing these figures. The first is that the maximum range of the Pegasus carrier aircraft is roughly 1000 nmi. This includes a captive carry to the launch site, an aborted launch, and a return to base with Pegasus still attached. The second assumption is that for launches that do not require the full advantage of Pegasus' mobility, the standard launch point for Pegasus out of the Eastern Range is 28°N , 281.5°E . The vertical axes in Figs. 9 and 10 represent the difference in argument of latitude between the two satellites (the angular separation within the same orbital plane). The horizontal axis represents the launch point as the difference in degrees from the nominal point listed before. The diagonal lines represent the difference in argument of latitude for each day in the first week of October, which was chosen simply as an example. Figure 9 represents the difference in argument of latitude for northerly launches (launch along the ascending pass). Figure 10 represents the difference in argument of latitude for southerly launches (launch along the

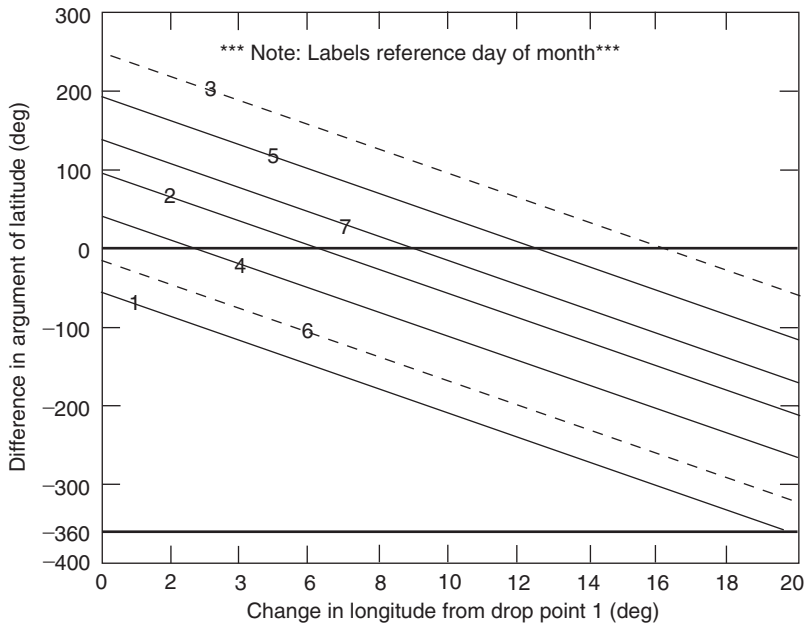


Figure 9. Graph of difference in argument of latitude for northerly launches.

descending pass). The horizontal lines simply demarcate zero angular separation between the two satellites.

The intersection of a diagonal line with a horizontal line defines a drop point within the range of the Pegasus carrier aircraft from which Pegasus can be

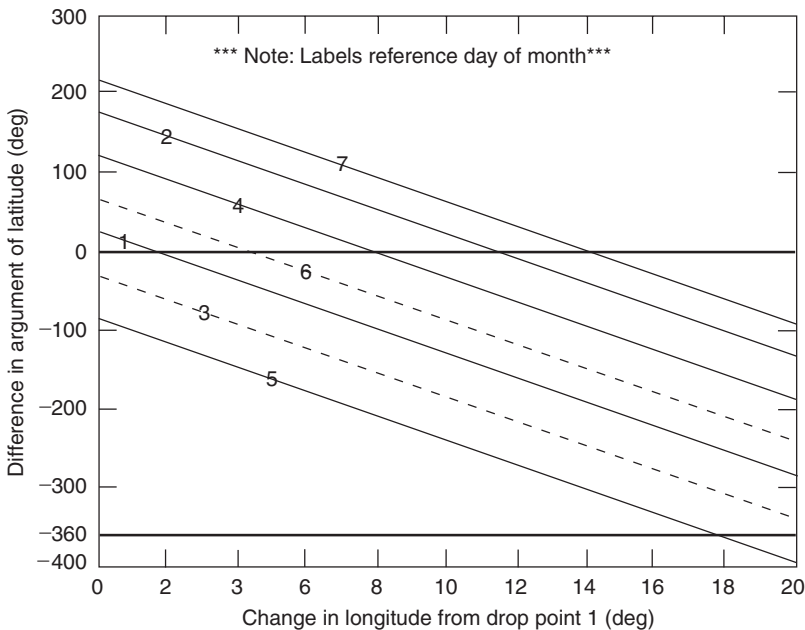


Figure 10. Graph of difference in argument of latitude for southerly launches.

launched and effectively deliver its satellite to the front door of the target satellite at the time of orbital insertion. Realistically, this is not how a rendezvous would normally be achieved. Ideally, the satellite being launched would be placed in a temporary parking orbit slightly below and behind the target satellite. Over the course of several orbits the distance separating the two objects would be slowly decreased using several controlled burns of the satellite just placed in orbit. This would imply that a drop point is needed not to achieve 0° difference in argument of latitude but some finite value. The example is still valid. Simply shift the horizontal lines up or down until the desired difference in argument of latitude is matched. Again, an intersection between a diagonal line and the horizontal line defines a launch point within the range of the Pegasus carrier aircraft that would result in the desired difference in argument of latitude.

As can be seen from Figs. 9 and 10, every day except two in the first week of October provides two launch opportunities. A southerly launch on the 3rd does not provide a drop point within the range of the carrier aircraft that will achieve the desired result. However, a drop point can be found on that day if the launch is along the ascending pass instead. Likewise, a suitable drop point cannot be found within the range of the carrier aircraft on the 6th of October when launching along the ascending pass, but one can be found if launching in a southerly direction. The same qualitative results would be obtained for any other time frame. The quantitative results might be slightly different. For instance, instead of having only one launch opportunity on the 3rd and 6th it may be the 4th and the 7th. But the end result is the same. The mobility of Pegasus and, by definition, Sea Launch, and Shtil provides ideal rendezvous launch opportunities at least once a day and in most cases twice a day.

Clearly there are disadvantages with all of these mobile assets. Pegasus is limited in its size due to the restrictions of the L-1011 and, more importantly, the mechanical limitations of the hooks that hold the vehicle to the plane. Sea Launch has somewhat of a temporal disadvantage in that it requires almost 2 weeks to travel to the launch site. Those problems are exacerbated for Shtil because its home port is farther north. Nonetheless, for some specific missions, the mobility and flexibility that are provided by these unique space-launched assets provide valuable supplemental services to the fleet of existing ground-launched vehicles.

BIBLIOGRAPHY

1. Powell, J. The China Lake Launches. *Air and Space*, pp. 367–378, Feb/Mar 1997.
2. <http://www.state.gov/www/global/arms/treaties/abm/abm2.html>.
3. *Pegasus Users Guide, Release 5.0*, Orbital Sciences Corporation, August 2000.
4. Isakowitz, S.J. *International Reference Guide to Space Launch Systems*, 3rd ed. AIAA, Washington, DC, 1999.
5. *Sea Launch User's Guide, rev. B*, Boeing Commercial Space Company, July 2000.

DALE FENN
Orbital Sciences Corporation
Dulles, Virginia

APOLLO 17 AND THE MOON

Apollo 17 was not the last flight of humans to the Moon. This writer was not the last human being to step on the lunar surface. More lunar exploration and even lunar settlement will occur, barring the future stagnation or disappearance of our civilization. Exploration and scientific investigations in the earth sciences are rarely complete, particularly for studies related to a specific field site. A long hiatus between field investigations may occur, but other forms of investigation, directly or indirectly related, continue. Apollo 17's field study of the Valley of Taurus-Littrow on the Moon in 1972 and subsequent examination of its significance to our understanding of the origin and evolution of that small planet and of our own constitute a good example of these facts of scientific life. As the third of the specifically "science" missions to the Moon in the twentieth century, Apollo 17 actually became the last lunar landing of the Apollo Program in September 1970 (1) rather than on 11 December 1972 when the mission reached Taurus-Littrow. The National Aeronautics and Space Administration (NASA) and the Administration of President Richard M. Nixon, with the acquiescence of the Congress, had concluded that no further planned amortization of the American taxpayer's investment in deep space exploration would be undertaken. As historically naive a political decision as this may seem today to some, it did not prevent the achievement of one of the Program's major goals—gaining a first-order understanding of the Moon and its relationships to the terrestrial planets. This became one of the primary historical legacies of the post-World War II generation.

Apollo had evolved quickly and radically toward increased scientific emphasis after Neil Armstrong first stepped on the Moon on 20 July 1969. Its purpose changed from the completed goal of meeting President John F. Kennedy's challenge (2) to land "men on the Moon and return them safely to Earth," to an objective of increasing human knowledge about the Moon and space. This would be done to the maximum extent possible using the technological and operational systems in hand and reasonable extensions of that capability. This shift in emphasis occurred smoothly and rapidly thanks to the foresight of senior NASA managers such as George M. Low, Apollo Spacecraft Program Manager; Robert Gilruth, Director of the Manned Spacecraft Center; Eugene Kranz, Chief of the Flight Control Division; Maxime Faget, Chief Engineer of the Manned Spacecraft Center; and General Samuel Phillips, Director of the Apollo Program. As early as the spring of 1969 (3), scientific packages were being enhanced, adding new experiments and improving old ones. Astronaut training in field geology, overseen by the author for the Astronaut Office, was altered to consist of field simulations (4) at geologically relevant sites using mission-specific equipment and procedures. These scientific training exercises taught pilots the art of geologic observation, sampling, and documentation and also put that learning in the context of real geologic problems related directly or indirectly to those they would encounter on the Moon. In addition to the U.S. Geological Survey's Principal Investigators for field geology, Eugene M. Shoemaker, Gordon A. Swann, and William R. Muehlberger (also of the University of Texas), world-renowned Earth scientists who doubled as outstanding teachers were given increased

access to mission planning, mission operations, and astronaut training. These new participants included Richard H. Jahns (Stanford University), Robert P. Sharp and Leon T. Silver (California Institute of Technology), James B. Thompson and James Hays (Harvard University), and Gene Simmons and William Brace (Massachusetts Institute of Technology). Also important to science was an increase in the capability of the Lunar Module (5) so later missions could include heavier scientific payloads, a lunar roving vehicle, and greater consumables, thus, longer time on the Moon. These augmentations meant a large increase both in time for exploration (22 hours for Apollo 17 versus a total of 19.4 hours for Apollo 11, 12, and 14, combined) and in distance traveled (35 km for Apollo 17 versus ~5 km combined walking distance for Apollo 11, 12, and 14).

The Apollo 11, 12, and 14 missions were flown largely under the original payload, training, and operational constraints imposed by the race to the Moon and the conservatism necessary for success in that race. These missions still managed, however, to produce remarkable suites of samples, photographs, and observations in addition to giving the Apollo team the operational confidence to land at more challenging but scientifically more interesting locations away from the lunar equator. In spite of the operational limitations, the analysis of samples and other information returned from the first three landing sites rapidly increased the understanding of the Moon and its history. Ironically, the Apollo 13 mission, which failed to land on the Moon, set the stage for the even more spectacular scientific returns from the last three landings, Apollo 15, 16, and 17. The crew and backup crew of Apollo 13 had embraced the new training emphasis on field geology and encouraged the Apollo 15 crew to follow suit. Apollo 13's backup crew, already convinced that a science focus was important, was assigned to fly the Apollo 16 mission. Finally, the designation of a scientist and geologist as the Lunar Module Pilot on Apollo 17 assured that all of the last three missions truly would be "The Great Voyages of Exploration" (6).

Due to the foreknowledge that Apollo 17 would be the last of the Apollo series, selection of its landing site became a contentious issue (7) among lunar scientists and between lunar scientists and operational planners. The usual candidates for landing sites reappeared: crater floor and central peak opportunities for deep sampling of the lunar crust, like the impact crater Copernicus; possible volcanic features, like the Davy Crater Chain and dark material in Alphonsus; and highland areas such as the rim of the crater Tycho and an area "Southwest of Crisium." Even a farside landing in the basin Tsiolkovskiy was given brief consideration due to the efforts of the author (8). Eventually, however, the scientists became increasingly interested in an unnamed, 2300-m deep, 50-km long valley, radial to the 740-km diameter circular basin, Serenitatis, that cut through the Taurus Mountain ring near the crater Littrow. This Valley of "Taurus-Littrow," however, was not a favorite of the operational mission planners. In spite of the pinpoint landing accuracy they had demonstrated on all previous missions since Apollo 11, the narrow valley, the mountainous approach, and the high valley walls gave the planners pause. Their legitimate concerns were compounded by the relatively short time, only 14 minutes, for navigational updates after acquisition of communications from the lunar module, Challenger, as its last orbit before landing carried it around the Moon from the farside. Initially, trajectory calculations indicated that three-sigma errors, the normal



Figure 1. Night launch of the Saturn V rocket carrying the Apollo 17 Mission to the Moon at 12:40 A.M., 7 December 1972 (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

extremely conservative planning limit, might result in hitting the side of the northern mountain wall. Gradually, however, refinements in navigational techniques for the mission and the inevitable synergistic give and take that so characterized Apollo interactions narrowed the three-sigma errors to about 1-km, the limit where all agreed that Taurus-Littrow could be the selected site. Thus, in late February 1972, only 9 months from launch, Taurus-Littrow was approved as the exploration site for Apollo 17 (9) (Fig. 1).

The Apollo 17 Mission

The Valley of Taurus-Littrow (Fig. 2) offered four major benefits as the last Apollo landing site, taken in the context of a final test of then current hypotheses related to the origin and evolution of the Moon. First, photogeologic analysis indicated that Taurus-Littrow provided access to a three-dimensional window into a mountain ring created by the Serenitatis large basin-forming event, by now well established as the result of a giant impact of an asteroid or comet. Second, major units of mare basalt and older nonmare rocks would be within easy reach of roving vehicle traverses. Third, a mantle of dark, possibly young volcanic debris partially covered the region as well as portions of the valley, and craters of a range of depths



Figure 2. The Valley of Taurus-Littrow as seen from the Commander's window on the left side of the Lunar Module Challenger on the orbit of the Moon before landing. The view is approximately west northwest, looking toward the Serenitatis Basin (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

penetrated this debris and the underlying basalt. And, fourth, the valley lies about 600-km north and 200-km east of the Apollo 11 and Apollo 15 sample areas, respectively, adding significantly to our exploration coverage of the Moon's nearside.

The Lunar Module crew of Apollo 17, Commander Eugene A. Cernan, and the writer as geologist and Lunar Module Pilot, conducted 22 hours of field exploration and experiment deployment in the Valley of Taurus-Littrow between 11 and 14 December 1972. During this period, the crew investigated, photographed, and sampled 11 major field locations. We traversed, observed, and sampled more than 35-km of the valley floor and obtained and documented 120-kg of samples from 97 major boulders and 75 other lunar materials. We took 2200 documentation photographs and deployed the 11 experiments of the Advanced Lunar Science Experiment Package (10). The crew had trained together for 15 months before launch, several days a month consisted of simulated traverses at field sites illustrating one or more of the types of geologic problems expected on the Moon and specifically at Taurus-Littrow. Combined with the geologic experience of the author, the organization and flexibility of the exploration plans (11–13), and the close cooperation of the science team in direct support on Earth, this training gave a stronger scientific foundation to Apollo 17's exploration that had been possible during previous Apollo missions (14,15).

The North and South Massifs constitute the major structural boundaries of the Valley of Taurus-Littrow. Slopes of an approximately constant 25° flank the Massifs, rising 2000 and 2300 m, respectively, above the valley floor. Discontinuous, but roughly horizontal exposures of thick sections of crustal rocks that predate the major mare basalt eruptions exist on the steepest upper slopes. These outcrops or near outcrops create numerous fields of exposed rock from which tracks lead downward to some of the sampled boulders at the base of the Massifs. Interlocking domes called the Sculptured Hills constitute Taurus-Littrow's northeastern wall and have concentrations of boulders apparent only on the inaccessible upper slopes of these hills. The valley floor consists of an undulating, highly cratered, relatively flat surface, covered largely by broken and pulverized basalt. One group of the cluster of craters surrounding the spot where Challenger landed, lies on a ray of secondary ejecta from the crater Tycho (16) 2000-km to the southwest. The largest of these craters is about 600-m in diameter. An older cluster of craters of about the same range of sizes cut into the floor northeast and to the west of the landing point, near the base of the North Massif. An irregular fan of material, the light mantle, projects northeast from the base of the South Massif. Finger-like projections of this fan reach out as much as 6-km from the Massif. Premission photographs suggested that a mantle of dark material covers the valley as a whole, including portions of the surrounding mountains. All surfaces are composed as pulverized debris called "regolith" (17), consisting largely of fragments of the bedrock below mixed with dark mantling material and other materials thrown into the area by more distant meteor impacts or introduced by volcanic eruptions younger than the bedrock.

Scientific activities in Taurus-Littrow (Fig. 3) began with the deployment of the experiments constituting the sixth and final Apollo Lunar Surface Experiments Package (ALSEP). This package had been enhanced to have a design life of 2 years rather than one (18). In connection with drilling holes for the heat flow experiment, two cores through the upper 3.2-m of the central valley regolith were obtained. Despite minor interruptions to work on technical problems with the ALSEP, about a third of the first excursion (extravehicular activity or EVA) and most of the second and third excursions concentrated on the planned traverses and exploration. Actual traverses followed this plan closely (19–22) except for a curtailed first traverse that only reached a point on the rim of Steno crater rather than reaching the original objective of Powell and deletion of the third traverse's Station 10 at Sherlock. Investigations of the basaltic mare materials south of the landing point began on the first excursion. On the second day, the traverse went west to sample premare materials at the base of the South Massif (Station 2). We then worked back over the light mantle deposit (Stations 2A and 3); to the dark, possibly volcanic crater, Shorty (Station 4); over the contact between light mantle and mare; and finally into the basalt boulder field surrounding the 100–150 m deep crater, Camelot (Station 5). The third day started with a long study of large boulders at the base of the North Massif (Stations 6 and 7), followed by sampling in the regolith at the base of the Sculptured Hills (Station 8), and a final stop at another possible volcanic crater, Van Serg (Station 9). Along each lunar rover traverse, we periodically sampled the surface of the regolith across various geologic units, deployed explosive charges for the active seismic profiling experiment, monitored the receiver for the surface electrical properties experiment,

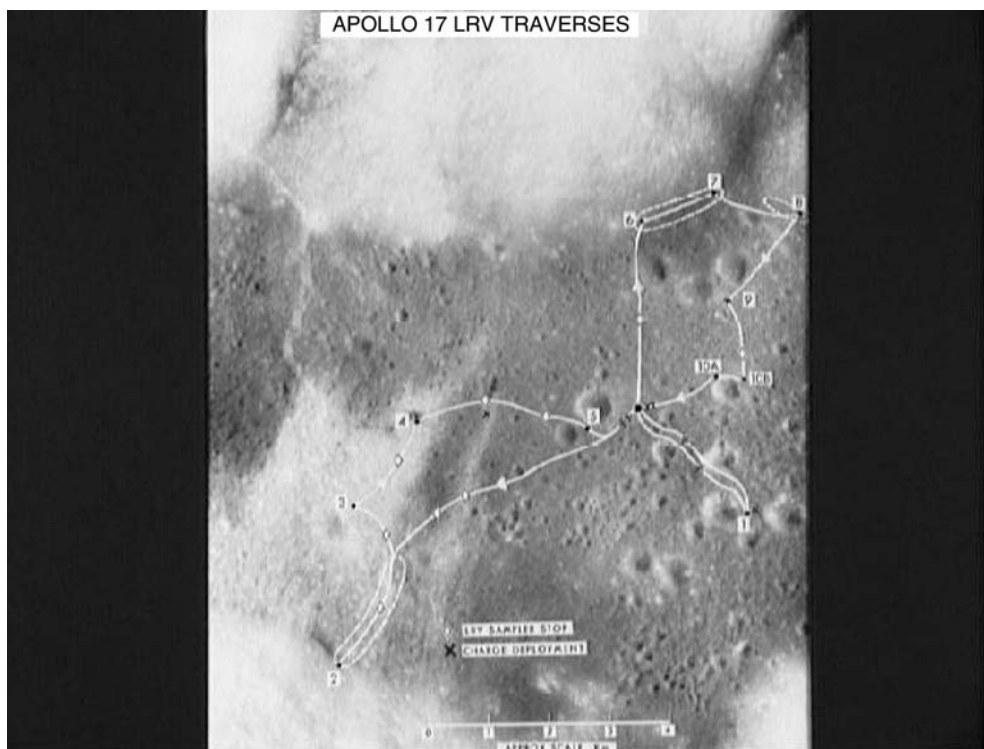


Figure 3. Apollo 17 exploration area in the Valley of Taurus-Littrow showing the landing site, exploration stations (numbers), and general traverses (solid lines) (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

and obtained readings from the traverse gravimeter (23). The principal sampling tools used included a rock hammer, a pair of long handled tongs, 35 and 70 cm core tubes, a long handled scoop (also used for trenching), and a supply of pre-numbered Teflon sample bags (Fig. 4).

Impact Cratering

Almost everything we think we know about the Moon must be viewed through the filter of impact cratering effects (Fig. 5) that have dominated lunar history from its origin to the present (24–29). The impact of comets, asteroids, meteors, micrometeors, dust, and energetic atomic and nuclear particles have modified the surface and near-surface expression of all of the internally generated processes that contributed to the present physical nature of the Moon. The secondary effects of each impact have magnified the importance of these impacts. Most comet, asteroid, meteor, and micrometeoroid impact velocities are between 13 and 18-km/s, and some are as high as 70-km/s, giving target pressures at the point of impact of several hundred Gpa (gigapascal). Extraordinary amounts of heat per unit mass are released as conversion of kinetic energy into forward and rearward shock waves takes place almost instantaneously. The amount of extralunar material that can be identified in regolith samples returned to Earth

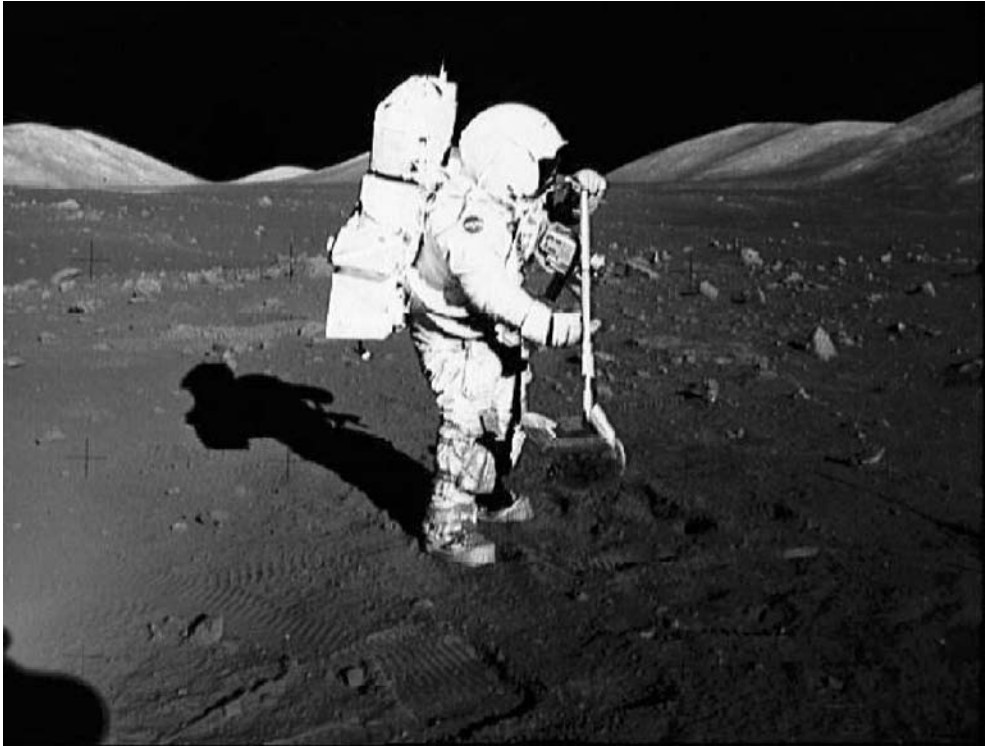


Figure 4. The author as a well-equipped astronaut on the Moon during the Apollo 17 mission in the Valley of Taurus-Littrow. He is using “the rake” sampling device to sift rock fragments from the finer portions of the regolith and has a 70-mm Hasselblad camera mounted on his chest (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

indicates that about 98% (30) to 99.7% (31) of all but the larger projectiles is melted, vaporized, or ionized, and returned to space. The general characteristics of lunar impact craters as a function of diameter are summarized in Table 1.

Processes associated with cratering and space radiation have created a well-defined zone of debris that covers essentially the entire Moon; its thickness depends on the length of exposure of a specific geologic unit or feature. This zone is called the “regolith,” a terrestrial term also used for the Moon. Essentially all the samples returned from the Moon by Apollo have come from the regolith or from rocks incorporated within it. It has been defined as “the layer or mantle of fragmental and unconsolidated rock material, whether residual or transported and of highly varied character, that nearly everywhere forms the surface of the land and overlies or covers bedrock. It includes rock debris of all kinds, including volcanic ash...lunar regolith consists [largely] of particles <1-cm in size although larger cobbles and boulders, some as much as several meters across, are commonly found...much of the pulverized material is melted and welded together to produce breccias (fragmental rocks) and impact melt rocks, which make up a significant portion of the regolith...”(32,33). A particularly important part of the lunar regolith consists of aggregates of rock, mineral, and glass fragments,



Figure 5. The 95-km diameter impact crater Copernicus as seen from the Apollo 17 Lunar Module Challenger after its departure from the Moon on 14 December 1972. Data from the Apollo 12 mission indicate that Copernicus formed about 900 m.y. ago (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

called agglutinates, held together by impact melt glass. Recently, it has been shown that on the nanometer scale, iron metal particles accreted on and formed in the rims of regolith grains significantly affect optical and magnetic properties (34–36). Further, the lunar regolith contains embedded solar wind gases, meteoritic material, and isotopic products and crystal structure damage produced by solar and cosmic radiation. The average depth of the regolith in a given area reflects the age of the underlying bedrock. Lateral mixing of material derived from adjoining bedrock units is a function of the age of the separating contact.

Origin and Evolution

A “standard” or “conventional” hypothesis for the origin and evolution of the Moon evolved significantly during the last three decades of the twentieth

Table 1. General Characteristics of Lunar Craters as a Function of Size^a

Diameter range examples	General characteristics
< 10-m	<ul style="list-style-type: none"> • Craters normally do not penetrate the regolith. • Depth to diameter ratio variable. • Glass discontinuously lines shallow pits in the center of fresh craters. • Mineral grains shattered around small craters on solid rock (zap pits). • Deep pits (~1/3 the crater diameter) in the center of some craters.
~ 10-m to ~100-m Van Serg and Shorty Craters in Taurus-Littrow	<ul style="list-style-type: none"> • Craters normally penetrate mare regolith if above 20-m diameter. • Depth to diameter ratio about 1:3 to 1:4 for fresh craters. • Inner benches common if target material stratified. • Regolith breccias present inside and on the ejecta blankets of young craters. • Ejecta blankets extend to about one crater diameter. • Target strata are overturned, but original vertical sequence is preserved in ejecta blanket.
~ 100-m to ~10-km Taruntius; Camelot Crater in Taurus-Littrow	<ul style="list-style-type: none"> • Both transient and initial steady-state craters are hemispherical and have circular and raised rims. • Depth to diameter ratio about 1:3 to 1:4 for fresh craters. • Impact breccias present inside and on the ejecta blankets of young craters. • Ejecta blankets extend to about one crater diameter. • Secondary impact cratering significantly modifies surface features out to many crater diameters from the edge of continuous ejecta. • Target strata are overturned, but their original vertical sequence is preserved in ejecta blanket.
~ 10-km to <300-km Copernicus	<ul style="list-style-type: none"> • Transient crater approaches hemisphere and has a circular raised rim and probably is lined with a shell of impact melt. • Initial steady-state crater has a flat floor and central mound or peak. • Initial steady-state crater walls have many stepwise benches (slump landslides) on walls. • Hummocky crater floors and the depressions on wall benches and near-rim ejecta blankets of larger craters have indications of pools and flows of impact melt. • Ejecta blankets extend to about one crater diameter. • Target strata are overturned, but their original vertical sequence is preserved in ejecta blanket. • Secondary impact craters, crater clusters, crater chains, and herringbone crater chains extend several thousand kilometers beyond edge of continuous ejecta.

Table 1. (Continued)

Diameter range examples	General characteristics
> 300-km (basin) Orientale	<ul style="list-style-type: none">• Transient crater depth to diameter ratio decreases with increasing size as lithostatic pressures compete with explosive pressures.• Transient crater has an increasingly flat trapezoidal crosssection and increasing diameters.• Transient crater has a flat floor, a circular raised rim, and probably is lined with a thick shell of impact melt.• Initial steady-state crater has a fractured, flat floor and central ring or partial ring of peaks.• Initial steady-state crater walls have many wide, stepwise benches (slump landslides) on walls.• Floors and depressions on wall benches and near-rim ejecta blankets have indications of large pools, mantles, and flows of impact melt. Impact melt also injected into target materials.• Ejecta and debris flow blankets extend beyond one crater diameter.• Two to six rings of mountains outside transient crater rim around basins > 400-km in diameter.• Target strata sequence is not well preserved in ejecta blanket due to extensive mixing of ejecta during flow.• Within one crater diameter of the final steady-state rim, there is a continuous deposit of melt breccia, possibly several hundred meters thick at the rim of the larger basins.• Secondary impact craters, crater clusters, crater chains, and herringbone crater chains extend beyond the edge of continuous ejecta and debris flows and reach thousands of kilometers and probably around the entire Moon to the basin antipode.

^aRef. 620.

century. This hypothesis currently holds that the Moon formed about 4.57 billion years (b.y.) ago by the aggregation of material produced during a giant impact between the very young Earth and a Mars-sized asteroid; most metallic core-forming material remained with Earth (37–43). Such an origin could explain the high angular momentum of the Earth–Moon system and at least some of the lunar geochemical constraints related to iron, volatile, and alkali elements other than potassium (44). Soon after or during lunar aggregation, lunar core formation occurred (45,46), and a Magma Ocean developed on its surface (47,48). The lunar Magma Ocean largely crystallized within 50 million years (m.y.) of the creation of the solar nebula and, at the same time, differentiated due to contrasts in mineral densities into an olivine-pyroxene dominated mantle and a 60–70-km thick Caplagioclase-rich crust (49,50). Late in this differentiation process, potassium, rare-earth elements, phosphorous, and thorium-rich residual liquid (urKREEP) (51) accumulated beneath the crust, largely in the region beneath what is now the Procellarum basin (52–56). Late ilmenite-rich cumulates (57) sank toward

the base of the less dense olivine and pyroxene cumulates carrying some urKREEP material with them (58). Intrusive and extrusive basaltic magmatic activity began soon after the Magma Ocean crystallized and, before the main sequence of mare basalt eruption began at about 3.8 b.y., produced the magnesium-rich suite of plutonic rocks (Mg-suite) (59–61), KREEP-rich basalts (62), and the cryptomaria (63–66). At about 3.85 b.y., a concentrated bombardment of the crust took place that produced most or all of the ~50 basins greater than 300-km in diameter visible today as well as most other observed cratering effects in the lunar crust (67–71). The effect of this late lunar “cataclysm” was to reset the ages of all crustal impact glasses yet studied (72,73). South Pole-Aitken basin is probably the only basin >1000-km in diameter to form during this late bombardment (74); the Procellarum basin is an artifact of the superposition of several smaller basins (75). A global magnetic field and presumably a circulating fluid metallic core were present at least between 3.9 and 3.8 b.y. (76–78). The core is now between 300 and 400-km in radius (79–81). Between 3.9 b.y. and about 1.0 b.y., mare basalts and basaltic pyroclastic materials erupted, largely on the nearside of the Moon (82–84). Major features on the Moon have been little modified subsequently other than the development of several meters of impact-generated regolith on most surfaces (85).

Although some aspects of this conventional hypothesis of lunar origin and evolution are attractive and probably correct, as will be discussed throughout this article, numerous difficulties exist in reconciling a number of its implications with everything we think we know about the Moon (86–89). Some of the major questions that can be raised with the conventional hypothesis are as follows:

1. Was the Moon formed as a result of a giant impact on Earth immediately after Earth’s accretion or was it formed independently and later captured?
2. Did core formation in the Moon and other terrestrial planets occur immediately after their accretion or was it delayed by the existence of a silicate protocore?
3. Did thermal convection and/or impact-induced splash cooling play a significant role in the crystallization and differentiation of the Magma Ocean?
4. Did the Moon’s Magma Ocean’s late ilmenite-rich cumulate sink near the base of the cumulate pile globally or only in response to local destabilization by the formation of a few extremely large impact basins?
5. Was the Moon’s Magma Ocean’s residual liquid (urKREEP) initially concentrated beneath the Procellarum Basin or distributed in a spherically uniform shell under the lunar crust?
6. Was the global thermal insulation effect of the impact-generated megaregolith of the lunar highlands critical to the later formation of the magmas that formed the basaltic maria?
7. Is the Procellarum Basin a consequence of the merging of several smaller basins or of a single extremely large impact?
8. Were the one or more extremely large impact basins and the ~50 large basins on the Moon the result of a “cataclysm” of impacts at about 3.85 b.y. or of a sustained bombardment lasting about 400 m.y.?

9. Was melting of the mantle due to pressure release (90) after large basin events significant in generating of magmas related to the Mg-suite of lunar rocks?

The scientific results of the Apollo 17 mission can now be viewed in the context of the conventional hypothesis and questions about that hypothesis from the perspective of more than 40 years of modern study of all of the Apollo missions and other lunar investigations. These later investigations have included telescopic and photogeologic mapping of the lunar surface, Apollo sample analyses, automated missions that both preceded (91) and followed Apollo, and the remarkable thought and computer modeling that has been stimulated by the collected data. Broadened multidisciplinary discussions of lunar origin and evolution are assisted by a descriptive formulation of the formative stages of lunar evolution as an augmentation of the traditional time-stratigraphic approach (92). The term “stage” is not used below in the normal time-stratigraphic sense (93). Rather “stage” is used in a more general sense for overlapping periods of lunar history that have definable but somewhat arbitrary beginnings and endings due in large part to the current incompleteness of information about the absolute ages of lunar events. Thus, the evolution of the Moon as a small planet (94–98) can be descriptively summarized as follows (Plate 1):

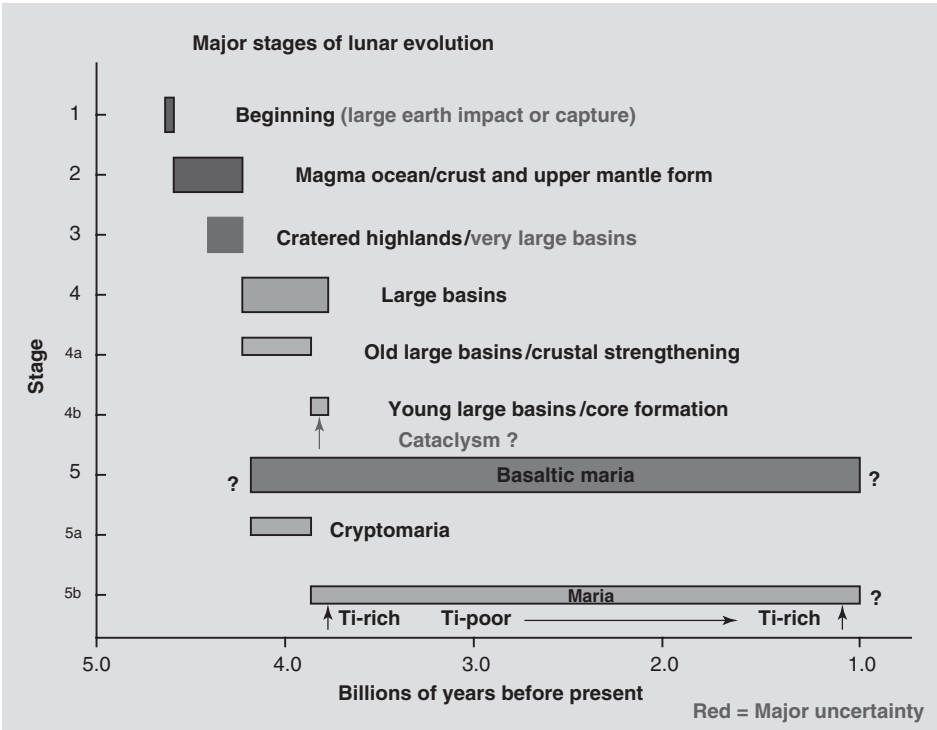


Plate 1. Major stages of lunar evolution. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

- Stage 1: Beginning [Pre-Nectarian (99)]—4.57 b.y. before present
- Stage 2: Magma Ocean (Pre-Nectarian)—4.57–~4.2(?) b.y.
- Stage 3: Cratered Highlands/Very Large Basins (Pre-Nectarian)—~4.4–~4.2(?) b.y.
- Stage 4: Large Basins—(Pre-Nectarian–Upper Imbrium)—~4.2(?)–3.8 b.y.
- Stage 4A: Old Large Basins/Crustal Strengthening (Pre-Nectarian)—~4.2(?)–3.92 b.y.
- Stage 4B: Young Large Basins (Nectarian–Lower Imbrium)—3.92–3.80 b.y.
- Stage 5: Basaltic Maria (Pre-Nectarian–Copernican?)—~4.2(?)–1.0(?) b.y.
- Stage 5A: Cryptomaria (Pre-Nectarian)—~4.2(?)–3.92 b.y.
- Stage 5B: Maria (Upper Imbrium–Copernican?)—~3.9–1.0(?) b.y.
- Stage 6: Mature Surface (Pre-Nectarian–Copernican)—~3.9 b.y.–Present.

Each of these formative stages overlapped significantly. The Magma Ocean began to form before the end of lunar accretion and probably was not fully solidified until after the end of the formation of old large basins. The Cratered Highlands overlapped at least the beginning of the Large Basin Stage. The Basaltic Maria magmas probably began forming initially by pressure-release (decompression) melting and then by thermal remelting of the upper mantle. Basaltic maria lavas first appeared on the lunar surface during the Large Basin Stage as cryptomaria and then partially filled many later large basins and other depressions. The regolith that underlies mature surfaces began forming on exposed units at the beginning of the Cratered Highlands Stage and continues to form today. Graphical cartoons illustrating this formulation of lunar evolution are referred to by Plate number in the following discussion.

Beginning (Stage 1). Discussion of the origin of the Moon (Plates 2 and 3) includes issues related to the origin of Earth and also to the origin of the solar system as well (100,101). One of the few undisputed scientific conclusions about the solar system as a whole is that it was formed from a concentration of interstellar dust and gas 4.567 b.y. ago. This conclusion is inferred from the radiometric ages of chondritic, eucritic, and iron meteorites (102–104) and from the initial isotopic ratios (radiometric model ages) of many lunar samples (105). Meteorites and lunar samples also preserve a record of extinct radionuclides. This record is consistent with the hypothesis that the formation of our solar system was initiated by an interstellar shock wave generated by a nearby supernova (106–108) that contributed the now extinct radionuclides and other materials to the solar nebula. The chemical similarity of carbonaceous chondrite (CI) meteorites to the composition of the Sun (109) and the current apparent abundance of such material in the solar system have led to the assumption that these meteorites closely represent the composition of primordial material that formed the Sun and the terrestrial planets. Computer models of processes in the early nebula have cast some light on what may have been happening after the shock wave during the first 10 million years or so (110–112). Once the initial angular momentum of the collapsing interstellar cloud had been dissipated and the rotating disk of the solar nebula had formed around the young Sun, particles began to stick together. This led gradually to the formation of planetesimals and then more rapidly to the

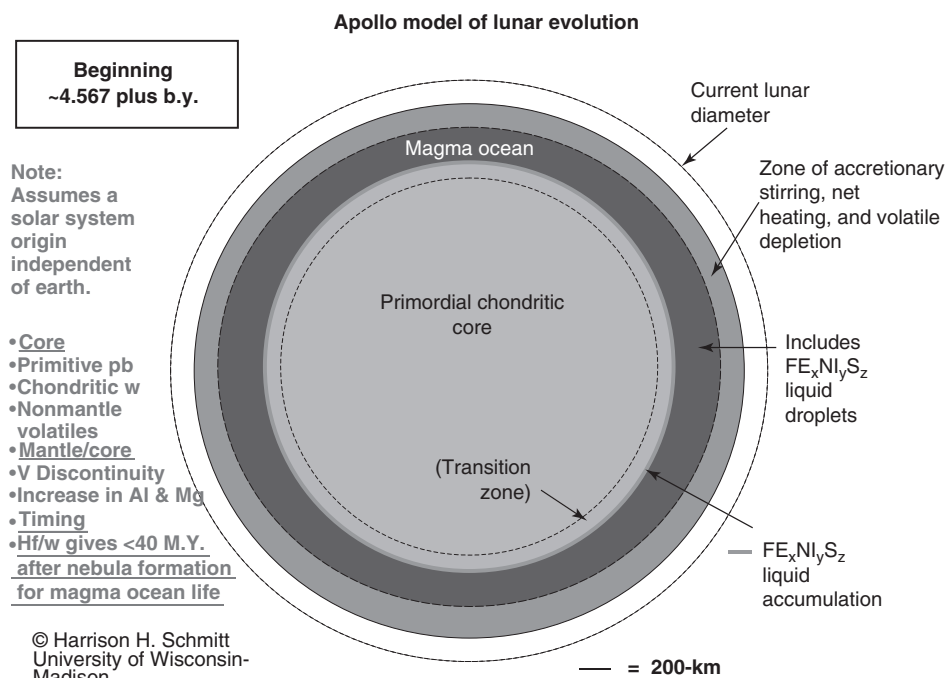


Plate 2. Apollo model of lunar evolution—Beginning ~4.567 plus b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

formation of planets. The relationship at this point between Earth and the Moon, if the Moon existed, remains unknown. Refractory element ratios (113,114) and oxygen isotope ratios (115–119) indicate, within the limits of analytical error, that the materials which accreted to form the two bodies came from the same accretionary region (feeding zone) of the solar nebula.

After the initial analysis of the results of the Apollo landings and subsequent robotic missions was completed, the question of the origin of the Moon boiled down to two competing hypotheses. As discussed before, a majority of present workers favors formation of the Moon as a result of the impact of a differentiated “Mars-sized asteroid” with a young but at least partially differentiated Earth. General computer modeling of such an origin suggests that it may explain the unusually high angular momentum of the Earth–Moon system (120,121). On the other hand, the amount of iron in the Moon (12%) in comparison to just Earth’s mantle (8%) appears to require that 90% or more of the Moon, formed in this “giant impact” model, must come from the impactor and relatively little from Earth’s mantle (122). A similar conclusion can be drawn from consideration of Hf/W systematics for Earth and the Moon (123). Thus, the giant impact model represents more of a Moon formed by impact-assisted capture of the mantle of the impactor rather than including a measurable fraction of the mantle of Earth. Increasingly, some of the geochemical inconsistencies of this computer model, particularly the evidence of an undifferentiated lower mantle discussed below, are being recognized (124–126). Other workers previously have suggested that the evidence supports the capture of the Moon by Earth after both

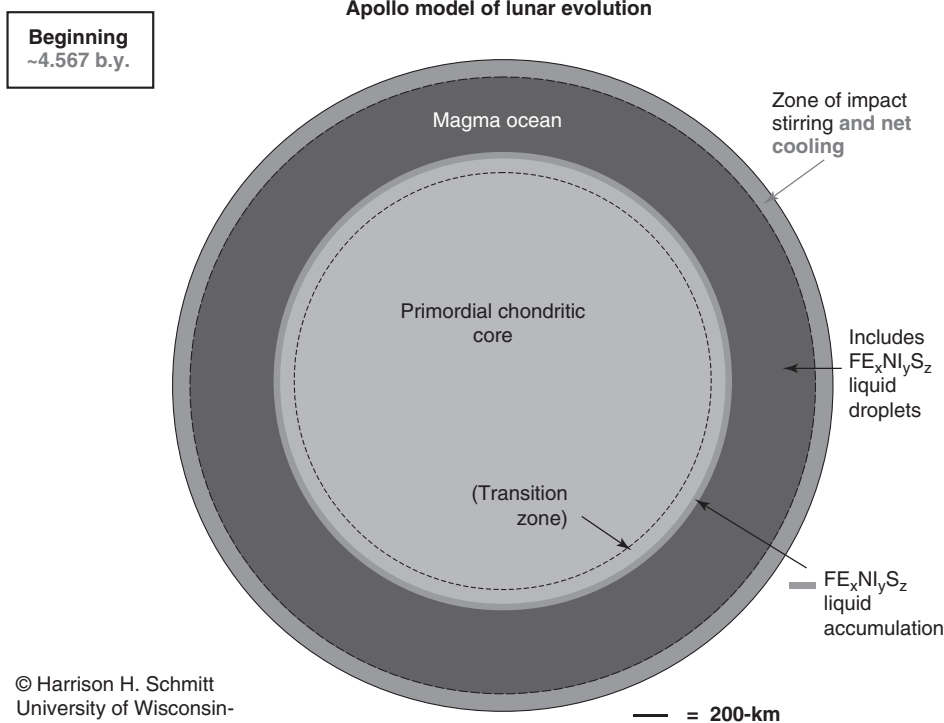


Plate 3. Apollo model of lunar evolution—Beginning ~4.567 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

formed independently as planets (127,128). In addition, models of giant impact scenarios are becoming increasingly confining in their possible initial conditions and subsequent evolutionary paths before the existing Earth–Moon system could have been created (129–136). Capture, provided that it is found compatible with the constraints of angular momentum, appears to be better than a giant impact in explaining the geochemical and geophysical structure of the lower mantle of the Moon. It has been suggested qualitatively, however, that special conditions related to a giant impact could produce the apparent undifferentiated characteristics of the lower lunar mantle (137,138). Considerations of lunar accretion from a circumterrestrial disk (139) also should receive additional attention in modeling studies (140).

No modern modeling studies of the capture hypothesis have been conducted since the early 1970s. Capture options, however, may exist that can explain the high angular momentum of the Earth–Moon system (141). Inherent in the capture hypothesis is the assumption that the Moon accreted in the inner solar system simultaneously with Earth, beginning with a cool, chondritic protocore upon which a Magma Ocean developed as the accumulated energy of accreting material increased with time. Gaining an early slight advantage in mass, co-orbiting Earth would have grown more rapidly and attracted proportionally more of the heavy element-rich planetesimals than the Moon, drawing from the same radial mixing orbital reservoir (142) in the solar nebula. The relatively

devolatilized character of the outer portions of the Moon, of which the Apollo samples are representative, would result from a much less competitive lunar gravity well relative to that of Earth both during accretion and during the impact-enhanced devolatilization of the lunar Magma Ocean. Alkali element depletions and isotopic ratios calculated for the Moon do not require the extreme devolatilization of a giant impact (143). Laboratory studies suggest that the volatility of sodium from a Magma Ocean is too low to account for the present proportion of sodium in the Moon (144). These studies, however, have not taken into account the pervasive stirring and spot high temperatures due to impacts that would enhance the efficiency of devolatilization of relatively volatile elements. Impact-generated temperatures of several thousand degrees would have existed locally near each impact point (145,146) and may have prevailed throughout the upper portions of the Magma Ocean during the last phases of extremely rapid lunar accretion. These portions of the Magma Ocean could be said to be boiling! Limits to devolatilization are indicated by the lack of fractionation between stable potassium and magnesium isotopes in the Moon; isotopic ratios of these elements are the same as Earth's (147,148). This lack of alkali isotopic fractionation and a rubidium/cesium ratio only a factor of about 2 greater than carbonaceous chondrites (149) indicate that the temperature of the impact-stirred portion of the upper Magma Ocean (although potentially several thousand degrees C) did not exceed that required for complete vaporization of alkali components. The reduced state of the Moon (150,151) also may relate to devolatilization of the Magma Ocean during which the ratio of water to iron was decreased to the point where all remaining water was disassociated by reaction with immiscible iron-rich liquid separating from the Magma Ocean.

The geochemical constraints of the lower mantle (below about 550 km), indicated by analyses of lunar pyroclastic glasses, also support the capture hypothesis. Volatiles and isotopic systems in the pyroclastic glasses that indicate largely undifferentiated source regions in the lower lunar mantle are incompatible with a Moon formed from the differentiated mantles of two impacting bodies, as required by the giant impact hypothesis of origin. Data from the Apollo 17 deposit of orange pyroclastic volcanic glasses have contributed significantly to this continuing debate over the origin of the Moon. The discovery of orange volcanic glass ("orange soil") in the rim of Shorty Crater (Station 4) and close to its original point of deposition (152), subsequent recognition of volcanic glasses as a widespread component of the lunar regolith, and detailed examination of these glasses in terrestrial laboratories established a number of new geochemical constraints on the origin of the Moon (153–155). Finding the orange glass (Figs. 6 and 7), whose geology is discussed under the Basaltic Maria Stage below, came from the convergence of a number of factors. First, one of the primary objectives of the mission to Taurus-Littrow was to search for the cause of dark mantling, possibly pyroclastic (explosive) volcanic deposits in the region. Second, premission consideration of multiple hypotheses for the origin of the dark-halo crater called Shorty included the possibility that it might be a volcanic vent, even though all evidence then available pointed to an impact origin. Finally, the writer's experience in volcanic provinces on Earth and his normal field geologist's instincts to watch one's feet probably were important factors. From the moment of discovery (156,157), it was clear that the orange glass would be significant, but



Figure 6. View of the Station 4 area, the site of the orange volcanic glass on the southern rim of the 110-m diameter Shorty Crater (to the right). The orange glass deposit lies just this side of the large boulder on the rim of Shorty, behind and to the right of the lunar roving vehicle and the author (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

how significant was not known until its return to Earth. It took detailed investigations in terrestrial laboratories before the ramifications of the glass slowly became clear. The primary importance of the pyroclastic glasses relative to lunar origin lies in the composition of the adsorbed volatiles on the surfaces of the small glass beads, and devitrified glass beads, as distinct from the glass itself. The volatiles washed from the surfaces of the beads are unlike any other lunar materials yet analyzed (158,159). A similar volatile component exists in the green volcanic glasses sampled by Apollo 15. Volatiles associated with both types of glasses are enriched over associated basalts by factors >100 in Cl, F, Br, Zn, Ge, Sc, Tl, and Ag and by factors >10 in Pb, Ga, Sb, Bi, In, Au, Ni, Se, Te, and Cu. These anomalies suggest that volatile components, otherwise depleted in lunar samples, are present in significantly higher abundances in the source areas of the volatiles accompanying pyroclastic glasses (160,161). The isotopic composition of the lead associated with the orange glass provides an even more unusual data point. It is isotopically primitive (162–166), that is, the lead had not been separated from its uranium and thorium radiogenic sources before eruption at the surface. Similarly, there is a chondritic initial tungsten value for the combined volatile and glass components, again suggesting that the lower mantle is largely undifferentiated and close to chondritic in composition (167).

Thus, the data related to volcanic glasses indicate that the source of the volatile components of the Apollo 17 orange glass lies below the devolatilized and differentiated cumulates of the Magma Ocean (see below) in a relatively



Figure 7. The Station 4 sample locality for the Apollo 17 orange volcanic glass, showing the dark, reddish orange central zone of the deposit and light-colored altered regolith at the left-hand contact (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

undifferentiated, possibly largely chondritic lower mantle. New assessments of Apollo seismic data suggest that the transition between the upper and lower mantle exists at about 550 km (168,169). A largely undifferentiated lower mantle does not fit the current model of lunar origin from debris generated by the impact of a differentiated planetesimal on a differentiated Earth. That debris would have lost much of its chondritic character prior to lunar accretion (170–172). Additionally, unlike Earth, the hafnium–tungsten isotopic system has significant variations within various categories of lunar samples (173,174) as a consequence of internal differentiation of these elements during the Magma Ocean Stage of lunar evolution (175). The mean model age (176) for the now extinct ^{182}Hf (half-life = 9 m.y.) and its stable daughter ^{182}W is about 50 m.y. in samples directly or

indirectly related to the differentiation of the majority of the lunar Magma Ocean (177). The orange glass isochron (178) also indicates an age of 54 ± 7.5 m.y. (179). These data are consistent with the idea that most of the Magma Ocean crystallized in 50 m.y. or less (180) and with theoretical analyses that Magma Oceans existed on terrestrial planets only very early in the history of the solar system (181).

Dynamical modeling and extinct isotope data (182,183) indicate that both accretion and planetary differentiation took place rapidly. A ~ 50 m.y. model age for forming major portions of the lunar mantle creates a very narrow, but not closed, window for a Mars-sized planetesimal and Earth to interact to form the Moon, considering that the total detectable life of ^{182}Hf would be ~ 90 m.y. Thus, within the first ~ 40 m.y. of ^{182}Hf 's existence, the following must occur:

1. The supernova that created ^{182}Hf and other now extinct isotopes must interact with and trigger the collapse of the interstellar gas and dust cloud to form the solar nebula.
2. Angular momentum within the solar nebula must be dissipated.
3. Planetesimals that have chondritic proto-cores must form in the inner solar nebula and aggregate into the terrestrial planets.
4. Magma oceans must form on some or all the terrestrial planets and be sufficiently stable that $\text{Fe}_x\text{Ni}_y\text{S}_z$ liquid can separate and accumulate at the base of the magma oceans.
5. Magma oceans must cool to a temperature where silicate crystallization and differentiation can begin.
6. $\text{Fe}_x\text{Ni}_y\text{S}_z$ metallic cores must displace the chondritic protocores of Earth and the impacting Mars-sized planetesimal if the Moon is to result from their impact interaction.
7. Sufficient silicate cumulates must form in the lunar magma ocean to isolate that portion of the Hf/W system in the source region of future mare basalts from the residual Magma Ocean.

Because steps 1–4 probably took 20–100 m.y. (184,185), then Hf/W systematics indicate that the formation of the Moon by a giant impact of another planet on the Earth must have occurred within a very narrow window of time—the first 30 m.y. or less. In spite of the narrowness of the opportunity, current modeling suggests that this is the most likely period for a giant impact to have occurred in the inner solar system (186,187). If such interaction did occur in this window and magma oceans existed on both objects at the time of impact (188), metallic core formation and initial mantle differentiation would have been incomplete. This would be inconsistent with what is known about the chemistry of the Moon and would require subsequent large accretionary impacts to adjust that chemistry to what is observed (189). Additionally, as discussed above, the apparent $^{182}\text{Hf}/^{182}\text{W}$ isochron age for the orange volcanic glass and its volatile component is 54 ± 7.5 m.y. This suggests that the deep source regions for the volatiles associated with the orange glass became closed systems for Hf and W when the volatiles' source areas in the more primitive lower mantle were no longer exposed to downwardly migrating $\text{Fe}_x\text{Ni}_y\text{S}_z$ liquid (see Magma Ocean discussion below).

Such migration would have kept the system temporarily open relative to tungsten. Thus, until the giant impact model for lunar origin or even an impact-assisted capture model or a circumterrestrial disk model can accommodate these geochemical and geophysical constraints, none of these models represent reality. At this time, the writer favors an origin by relatively passive Earth capture (190); however, most workers in lunar science favor origin by giant impact. No evidence has been recognized that would suggest when capture might have occurred; however, apparent lunar tidal signatures in 2.0 b.y. (191) and 3.2 b.y. (192) old terrestrial sedimentary rocks indicate a lunar influence on Earth's oceans by at least that time. The 3.2 b.y. tidal signature also has been interpreted to indicate that the lunar orbital shape at that point was similar to the current shape. If tidal interaction between Earth and the Moon's iron-rich core were necessary to produce a dynamo-generated lunar magnetic field (193), then the Moon had been captured by ~ 3.85 b.y. ago. This is the age of the large impact basin Imbrium (194), the oldest large basin that has a clearly identified remnant magnetic field (195) antipodal to it.

Magma Ocean (Stage 2). A consensus has developed from the data from all of the Apollo missions that, as accretion of the Moon accelerated, a silicate Magma Ocean was created (196–201) (Plate 2). Although debate continues as to the depth of the lunar Magma Ocean, the geophysical constraints (202–204) appear to indicate that only the outer ~ 500 -km melted, another 50–100-km partially melted, and a “primitive accretion core” (205), of roughly 1200-km radius, remained relatively cool and solid. Melting of the accreting material was accomplished by the conversion of accretionary kinetic and potential energy to heat and by the decay of short-lived radioisotopes such as aluminum-26 (206). As the Magma Ocean formed, the dynamic heating, splashing, and stirring of the global melt would have delayed differentiation (relative separation of the elements) due to fractional crystallization, a process by which minerals more dense and less dense than the parent melt sink or float, respectively, as they crystallize. As discussed before, some volatile differentiation, due to the thermal escape of elements too light and thermally energetic to be held by lunar gravity, would occur during this formative period, provided that the temperature of the Magma Ocean exposed to space reached several thousand degrees C. Thus, water, carbon dioxide, nitrogen, noble gases, and highly volatile elements, such as bismuth (Bi), thallium (Tl), indium (In), and cadmium (Cd) (207), are strongly depleted in the outer Moon relative to solar and carbonaceous chondritic abundances (208). Moderately volatile elements, such as lithium (Li), sodium (Na), potassium (K), rubidium (Rb), and cesium (Cs), are somewhat depleted as well (209). Sodium is depleted by a factor of 3–5 compared to Earth (210). The rare-earth elements, however, do not show evidence of selective loss based on their range of volatility (211). This suggests that at least some of the apparent elemental depletion patterns in lunar materials were inherited from the materials from which it was formed (212,213), either the differentiated mantle of a large impactor (giant impact), a circumterrestrial postaccretion disk, or differentiated planetesimals elsewhere in the inner solar system (capture). Also, as the materials from which the Moon formed probably consisted of largely oxidized phases that had chondritic affinities, initial separation of liquid metallic iron may not have occurred in the early Magma Ocean. Rather, as accreting material began to melt, iron and

nickel, along with sulfur, would be concentrated in an immiscible (214) metal-rich liquid ($\text{Fe}_x\text{Ni}_y\text{S}_z$) (215,216) accompanied by some portion of other siderophilic elements (217), including cobalt, precious metals, and platinum group elements (PGE). The actual cobalt and nickel contents for the Moon as a whole and for the lunar mantle in particular are still uncertain (218,219) and cannot yet be relied upon to help unravel the question of lunar origin. An iron and sulfur-rich liquid would have been an immiscible in the predominantly silicate liquid of the Magma Ocean. As iron-rich liquids are very dense, $>4.5\text{-gm/cm}^3$, a rain of small droplets would aggregate rapidly on the floor of the Magma Ocean (density $\sim 3.3\text{-gm/cm}^3$) and ultimately move toward the center of the Moon through the cooler lower mantle. In the process of nucleation in the Magma Ocean, the immiscible $\text{Fe}_x\text{Ni}_y\text{S}_z$ droplets would react with any remaining water to form hydrogen and iron oxide (wustite). Hydrogen would be dissolved in the silicate liquid, maintaining a reducing environment in the Magma Ocean early in its differentiation history. That this occurred is consistent with the necessarily reduced state of europium ions in the Magma Ocean during differentiation (220) and with experimental work (221) relative to chromium abundances in volcanic glasses. Both facts indicate low oxygen fugacities (measures of the tendency toward oxidation) in the Magma Ocean.

The next major change in the Magma Ocean began with the sequential crystallization of silicate minerals (222–224) and their separation from the liquid in relation to contrasting densities, that is, fractional crystallization (Plate 4). This critical chemical differentiation process started when the rate of heat addition due to accretion and radioactive decay became less than the rate of heat loss, and the temperature of at least portions of the Magma Ocean decreased to about 1300°C . Minerals forming in a silicate liquid have compositions different from that liquid, so chemical change in the liquid, that is, differentiation, results when those minerals are separated from the liquid. Assuming that the composition of the Magma Ocean resembled an iron, nickel, sulfur, and volatile-depleted carbonaceous chondrite, the first silicate mineral to begin to crystallize was probably magnesium-rich olivine [$(\text{Mg},\text{Fe})_2\text{SiO}_4$], followed closely by magnesium-rich orthopyroxene [$(\text{Mg},\text{Fe})\text{SiO}_3$] (225) (Plate 4). Olivine and orthopyroxene crystals below the zone of likely continued impact-induced turbulence sank to the floor of the Magma Ocean as cumulates. There, they were immersed in the ocean's residual silicate liquid which ultimately solidified interstitially, forming more olivine and pyroxene but eventually crystallizing calcium-rich plagioclase [$(\text{Ca},\text{Na})\text{Al}(\text{Al},\text{Si})\text{Si}_2\text{O}_8$], ilmenite [FeTiO_3], and other minor minerals. These interstitial minerals make up only a few percent of the final cumulate mantle rock but included trace amounts of radioactive isotopes of potassium, uranium, and thorium, important in the later remelting of the upper mantle to generate the mare basalts (see mare basalt discussion below). Above the roughly 500-km deep presumed base of the cumulate zone, seismic velocities (226) are consistent with those measured for a mixture of magnesium-rich olivine (75–85% forsterite component) and pyroxene. Mare basalt melting experiments suggest that pyroxene in the upper mantle probably varies upward from orthopyroxene dominant to pigeonite [$(\text{Ca},\text{Mg},\text{Fe})(\text{Mg},\text{Fe})\text{Si}_2\text{O}_6$] dominant to calcium clinopyroxene [$\text{Ca}(\text{Mg},\text{Fe})\text{Si}_2\text{O}_6$] dominant. These minerals and their compositions are consistent with phases found in equilibrium with the mare basalt magmas at

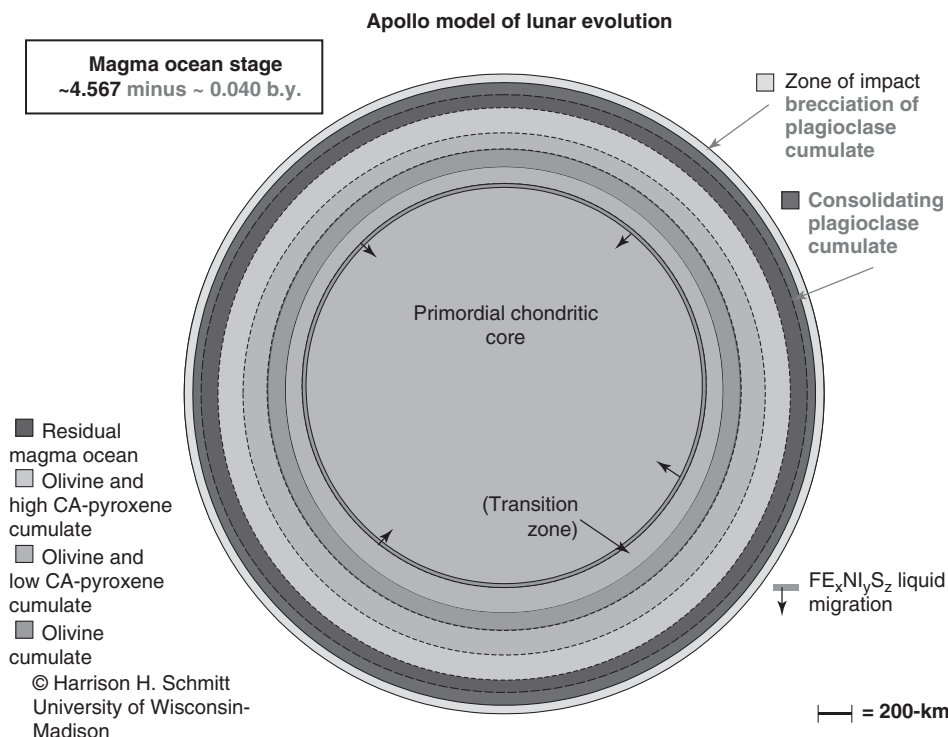


Plate 4. Apollo model of lunar evolution—Magma Ocean Stage ~4.567 minus ~0.040 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

pressures corresponding to their apparent depths of origin (227). The consistent relation between seismic data and the experimental work on mare basalt melting suggests that the combined effects of convective circulation (228) in the crystallizing Magma Ocean were minimal. In fact, impact-related mixing and splashing plus conduction of heat probably contributed more to early cooling than convection. Thus, early crystallization may have taken place in the upper, more rapidly cooled portions of the Magma Ocean, causing incorporation of relatively undifferentiated Magma Ocean liquid as interstitial material in early cumulates.

Crystallization of calcium-rich plagioclase feldspar, anorthite [(Ca,Na)(Al,Si)Si₂O₈], joined the formation of olivine and pyroxene in the Magma Ocean after sufficiently large volumes of such cumulates had separated to raise the aluminum concentration in the remaining liquid to the level required for plagioclase nucleation (229). Assuming chondritic concentrations of aluminum (<5% Al₂O₃), this point of plagioclase saturation in the residual Magma Ocean would have been reached after about 80% of the original magma had crystallized or Al₂O₃ concentration had reached about 15% (230). One consequence of plagioclase separation from the Magma Ocean was depletion of divalent europium relative to other rare-earth elements in the residual magma; the evidence for this was apparent in the first analyses of lunar basalts (231,232) and lunar anorthosites (233). Europium depletion in basalts and concentration in anorthosites constitutes one of the major pieces of evidence that there had been a lunar Magma

Ocean (234). Slight depletion of europium in pyroclastic orange and green glasses (235,236), however, and the evidence that these glasses were partial melts of deep (early) cumulates (237,238) indicate that there may have been some convective interchange with the residual Magma Ocean during the period of early plagioclase separation. The early formed plagioclase had probably less than 2–3% (239,240) of the albite component $[\text{NaAlSi}_2\text{O}_8]$ in solid solution. The albite component would be expected to increase gradually with time to more than 25% in plagioclase (241) forming from the Magma Oceans' residual liquid, as sodium and aluminum concentrations in the remaining liquid increased, relative to calcium, and olivine and pyroxene continued to separate. So far, examples of relatively sodium-rich ferroan anorthosite are rare, and those that have been found in the sample collections are less iron-rich than would be expected (242,243). Plagioclase crystals (density 2.9-gm/cm^3) would tend to float toward the top of the Magma Ocean; however, if there were continued impacts of debris into the Magma Ocean and some downward convective currents, their aggregation into a coherent lunar crust would have been slowed significantly. Plagioclase-rich samples from Apollo 15, 16, and 17 highland breccias also contain pigeonite (an iron-bearing clinopyroxene) and appear to represent an impact-modified version of this primitive crust (244). The detailed evolution of ferroan anorthosite crustal rocks and their mineralogical and geochemical variability remain unclear (245).

The overall crust, commonly referred to as "ferroan anorthosite," contains about 85% plagioclase by volume, and the rest is mostly pigeonite (246). Recent examination of feldspathic lunar meteorites and their comparison with Clementine optical remote sensing data suggest that the uppermost crust may be about 83% plagioclase and remainder is about 4.2% FeO, 5% MgO, and less than 0.5% TiO_2 (247). Apollo seismic data and gravity data from Clementine orbital perturbations (248,249) indicate that the anorthositic crust has an average thickness of about 60–70-km. Thickness ranges from a maximum of about 120-km in the farside region between the giant Procellarum and South-Pole Aitken Basins to a minimum of about 20-km in several of the young, large basins on the nearside. Such extreme variations, although model dependent and constrained primarily by Apollo nearside seismic data, may be largely if not entirely the result of the globally asymmetrical redistribution of crustal material by large basin-forming events (see below). It can be argued that the term ferroan anorthosite should be modified to ferroan "gabbroic or noritic" anorthosite (250) as a more appropriate designation for the upper crustal composition, given the content of less than 90% plagioclase (251). There may be another important type of lunar crustal rock, referred to as "granulitic breccia" (252) and identified in Apollo 16 and 17 samples (253–256). In fact, these rocks could be called "magnesian gabbroic anorthosite" as they have $\text{Mg}/(\text{Mg} + \text{Fe})$ ratios up to 75 or five points greater than the <70 of the ferroan anorthosites. These rocks may have genetic affinities with the Mg-suite of rocks (see below), although that is not clearly apparent at this time. Continued refinement of the relative distribution of magnesium and iron in the crust through Lunar Prospector and Clementine data synthesis may ultimately provide a better definition of global relationships involving magnesium granulitic breccias (257). In summary, a synthesis of current data and interpretations suggest that there are three major zones in the lunar crust. The upper 10–30-km is ferroan

gabbroic anorthosite that represents an original ferroan anorthosite with plagioclase >95% contaminated by (1) early splash from impacts penetrating into the residual Magma Ocean and/or (2) more pyroxene-rich material excavated from the lowest zone of the crust by large impacts. The middle 10–20-km is ferroan anorthosite that represents a largely unmodified plagioclase floating cumulate from the Magma Ocean (258). The lower 20–100-km is ferroan anorthosite that contains numerous Mg-suite intrusions (259) possibly associated with magnesium anorthosite, the so-called granulitic breccias.

Analyses of samples of Apollo 17 rocks consisting of fragments of Mg-suite rocks, or breccias, at Stations 2, 6, and 7, have provided new, but complicated insights into processes related to the lunar Magma Ocean and possibly to the formation of the lunar crust. A particular class of nonmare rock fragments that contain magnesium-rich mafic minerals and calcium-rich plagioclase exist in these breccias (260) and are composed of only one to three principal minerals, namely, olivine (dunites) (Fig. 8), pyroxene (pyroxenites), plagioclase and orthopyroxene (norite), and plagioclase and olivine (troctolites). These rocks, and similar rocks sampled at other Apollo sites, are known as the “Mg-suite.” Some rocks composed largely of plagioclase (anorthosites but not ferroan anorthosites) and possibly some alkali-rich rocks (261), appear to be related to this group. Although this is a very complex suite of rocks and subject to much debate, they probably are at least the result of the relatively undisturbed fractional crystallization of magnesium and calcium-rich silicate magmas, possibly of a variety of initial origins, that solidified in large chambers within the early crust. Multiple origins for the Mg-suite are suggested by their contrasting chemical characteristics (262) as well as their range of ages. Early remelting of Magma Ocean cumulates due to impact-induced, pressure-release (decompression) melting caused by cumulate overturn and/or impact removal of crust may have resulted in the formation of the oldest Mg-suite magmas (263) (Plate 5). Impact-induced remelting of magnesium-rich cumulates is consistent with the 4.4–4.5 b.y. crystallization ages (264) of a dunite (olivine) from Station 2 and a troctolite (plagioclase + olivine) from Station 8. Alternatively, the Mg-suite may be related to the materials that constitute the Procellarum KREEP Terrain (265–270) and not a broadly distributed, intrusive component of the lunar crust. The geographical limitation of Apollo and Luna samples to eight sites in a relatively small portion of the lunar nearside prevents full evaluation of this suggestion. On the other hand, the global distribution of apparent Mg-suite materials exposed by relatively young impact craters in the crust suggests no such geographical limitation (271). Another suggestion (272) that some of the older examples of the Mg-suite may be the consequence of impact induced “splash intrusions” of unsolidified Magma Ocean is not compatible with the increasingly higher iron content that would be present in such residual Magma Ocean liquids. The chemical characteristics of the younger examples of the Mg-suite of rocks, (>3.9 and <4.4 b.y.) also may indicate complex scenarios of partial melting of selected portions of the upper mantle, assimilation of crustal materials, and subsequent fractional crystallization in the crust (273,274). Recent work (275) has suggested that the source materials for some of the Mg-suite that contain particularly Mg-rich olivine are zones of intermediate depth in the mantle, rich in olivine and calcium-poor pyroxene. These cumulate zones, never very cool and

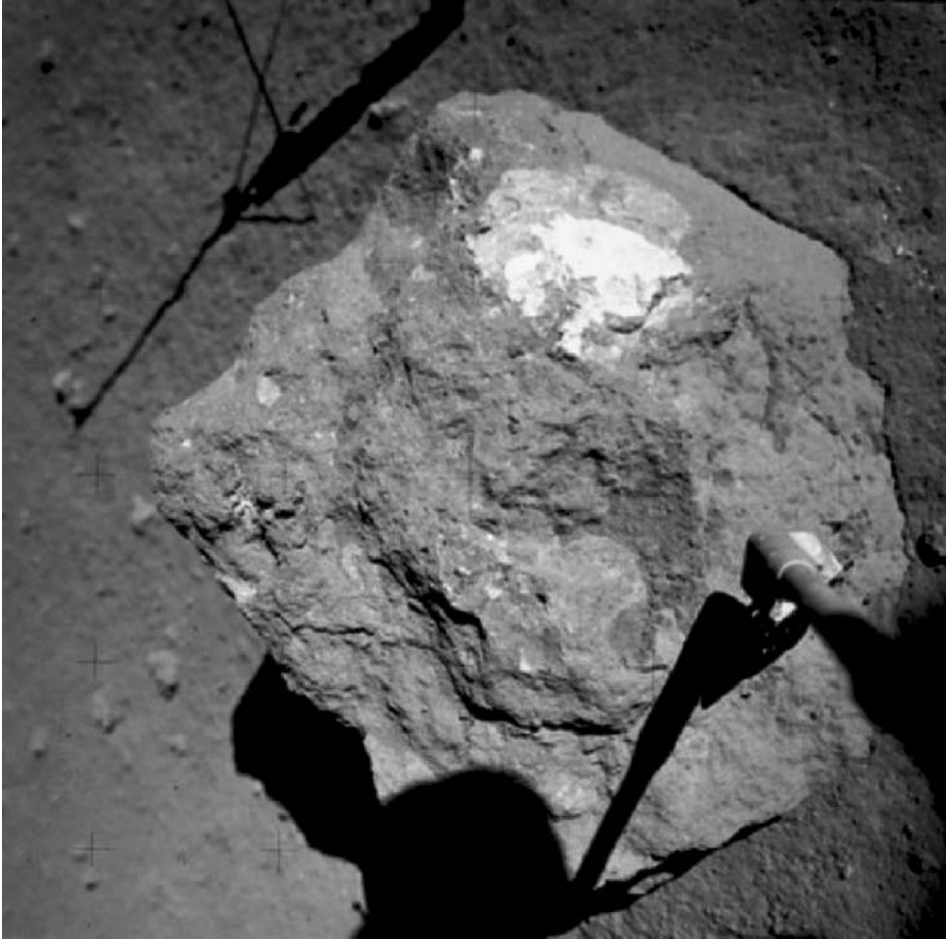


Figure 8. The clast of nearly pure olivine rock (dunite) sampled by the Apollo 17 crew at Station 2 at the base of the South Massif and dated as having crystallized 4.5 b.y. ago. The clast is enclosed in a blue-gray impact breccia and is a member of the Mg-suite of lunar rocks (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

beginning to reheat under an increasingly insulating crust (see the Cratered Highlands Stage below), may have partially melted due to the release of pressure during the formation of the earliest of the old, large basins (see below). Intrusion of such pressure-release melts into the crust, assimilation of some crustal debris, and subsequent fractional crystallization may explain the younger examples of the Mg-suite. The restricted mineralogical makeup of the rocks of the Mg-suite is consistent with initial solidification as fractionally crystallized, probably layered bodies. These layered intrusives would have been broken up and excavated by repeated large impacts until clasts in the breccias are all that remain to be sampled at the lunar surface. Mg-suite intrusions, however, may remain as coherent bodies in the lower crust, as evidenced by masses of apparently related material identified by remote sensing in the central peaks, walls, and ejecta

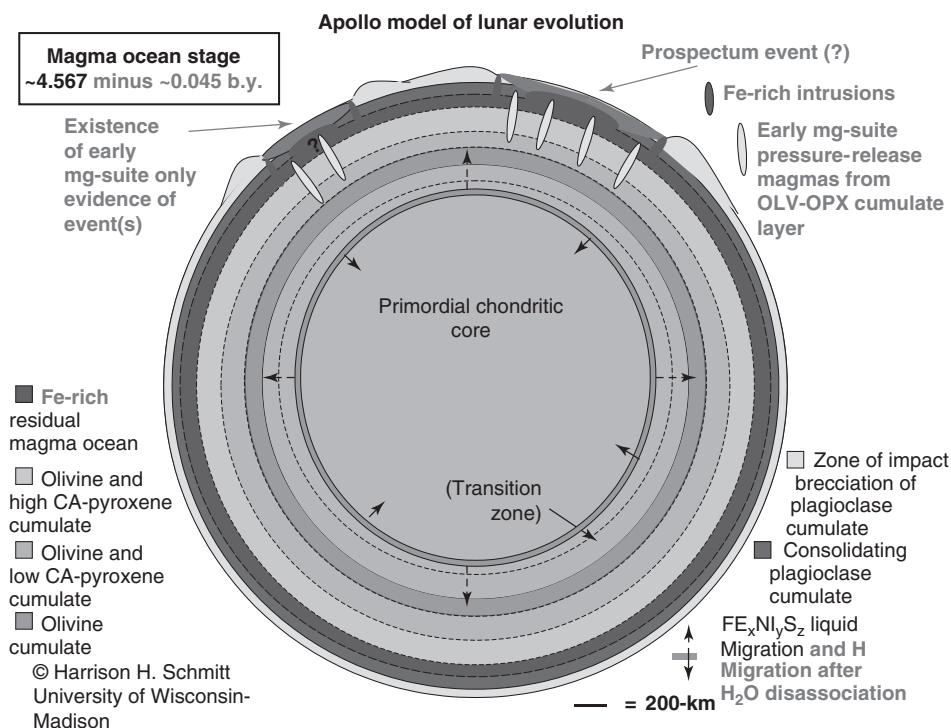


Plate 5. Apollo model of lunar evolution—Magma Ocean Stage ~4.567 minus ~0.045 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

blankets of many impact craters (276–278) and by the increase in seismic velocities in the lower crust (279). Many characteristics of these remotely sensed masses, such as their magnesium to iron ratios, have not yet been determined, so it cannot be conclusively said that some or all are related to the Mg-suite, but rather may be related to the ferroan anorthosite or magnesium granulitic breccia magma systems (280).

When the crystallization of the Magma Ocean was about 90–95% complete (281,282) and as iron and titanium concentrations in the residual Magma Ocean increased, ilmenite [$FeTiO_3$] became the last major mineral to crystallize and settle from the residual magma (Plate 6). Ilmenite-rich cumulates (3–12% ilmenite), accompanied by continuing formation of increasingly iron-rich olivine and pyroxene and increasingly sodium-rich calcium plagioclase, became an important constituent of the upper portion of the lunar mantle. For example, high-titanium mare basalt magmas produced by later partial melting of the mantle appear to have originated at depths between 100 and 200-km below the surface (283), plausibly a region of ilmenite-rich cumulates. The trace element ratios of such basalts also support their origin from partial melting of late ilmenite-rich cumulates. Europium depletion in the titanium-rich basalts, as in all mare basalts (284–286), indicates extensive previous plagioclase separation from the parent Magma Ocean. Further, ilmenite's precipitation from the Magma Ocean would deplete niobium in the residual liquid (287). High Nb/U and Nb/Th

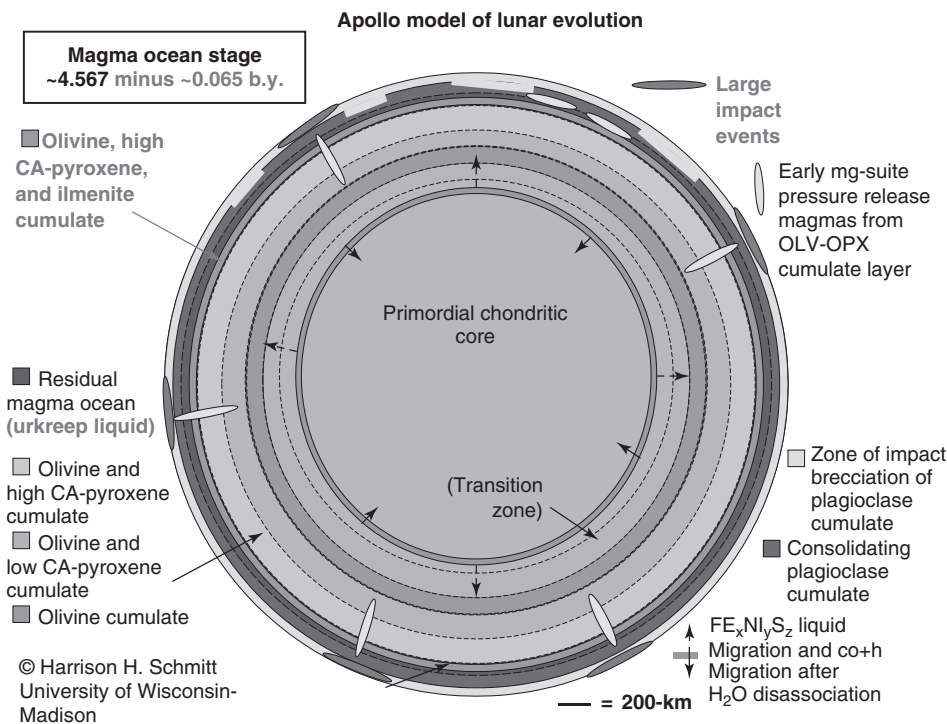


Plate 6. Apollo model of lunar evolution—Magma Ocean Stage ~4.567 minus ~0.065 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

ratios in high-titanium Apollo 17 basalts (288) reflect the consequent enrichment of this element in the ilmenite-rich cumulates and are consistent with an origin for these basalts through partial melting of such cumulates. Ilmenite crystallization also depleted hafnium in the remaining liquid (289). The high-titanium basalts sampled by Apollo 17 (see discussion of the Basaltic Maria Stage, below) have high radiogenic ^{182}W and high Hf/W ratios (290). This is consistent with ilmenite cumulate crystallization while significant ^{182}Hf (half-life 9 m.y.) existed, probably less than ~50 m.y. after the origin of the solar system (291). There have been recent suggestions that the late ilmenite-rich cumulates, once formed, sank to greater depths in the other cumulates of the upper mantle (292–294) because of ilmenite's high density relative to the underlying olivine–pyroxene cumulates (4.7 gm/cm^3 vs. $\sim 3.5 \text{ gm/cm}^3$, respectively). The average density of the ilmenite cumulate, however, would be only slightly higher than the olivine–pyroxene combination. The apparent relatively shallow depth of origin of titanium-rich basalts sampled by Apollos 11 and 17 would also contradict this hypothesis, at least for the eastern regions of the lunar nearside.

When ilmenite appeared, the continued crystallization of the four major minerals produced by the Magma Ocean (olivine, pyroxene, plagioclase, and ilmenite) largely completed the formation of the upper lunar mantle (that is, the mantle above about 500-km and below the base of the crust). After about 99% of the Magma Ocean had solidified (295), however, a geochemically significant

amount of residual silicate liquid would have remained below the crust. Fractional crystallization of the major minerals would have enriched this liquid in silicon (Si), potassium (K), and incompatible elements (296), that is, most rare-earth elements (REE), phosphorus (P), uranium (U), and thorium (Th). Dissolved volatiles, including hydrogen (H) and carbon monoxide (CO), also would be concentrated in this remaining magma. This residuum has been named “urKREEP” to distinguish it from an apparently related component in Apollo samples that was originally called just “KREEP.” Tungsten (W) also would be enriched in urKREEP and represents a special case. After the early separation of much of the siderophilic trace elements by the rapid precipitation of $\text{Fe}_x\text{Ni}_y\text{S}_z$ liquid, any residual such elements, including tungsten and subsequently formed radiogenic ^{182}W , would be gradually concentrated in the residual Magma Ocean as incompatible elements. As a residual silica-rich liquid, urKREEP would have existed as a liquid at temperatures lower than those of the original Magma Ocean. Such a liquid would have potentially stayed molten as a result of the concentration of radioactive, heat-producing isotopes and the insulating effects of the Cratered Highlands forming on the upper crust (see discussion below).

Some outward transfer of heat by convection in the cooling Magma Ocean seems likely during much of the Magma Ocean Stage (297–299). On the other hand, as solidification proceeded, potential convection cells would have been increasingly restricted. Convective tendencies also would have been reduced by cooling of the young Magma Ocean through continued impact splashing in its upper portions, a major factor not yet included in models of thermal evolution. The importance of splash cooling may be suggested by indications, discussed before, that the Magma Ocean largely solidified in 50 m.y. or less. Subsequently and throughout lunar history, there is no direct evidence that mantle convection was a significant thermal transfer process. That convection was not important is indicated by recent thermal modeling investigations (300). Further, little or no evidence of mantle convection exists at the surface unless the $\sim 2000\text{-km}$ long, north–south volcanic ridge system in the western portion of the Procellarum Basin implies that such activity existed beneath this region (301). (Alternatively, this Procellarum ridge system may reflect a cryptic crustal structure associated with the transient crater of the Procellarum basin-forming event discussed below.) After the Magma Ocean crystallized to become the upper mantle, conduction and upward migration of partial melts were probably sufficient to prevent any major density instabilities that would lead to solid or nearly solid-state convection in the one-sixth gravity environment of the Moon. Cooling sufficient for the solidification of all but the radioisotope-rich, late residual liquid of the Magma Ocean took place in about 150 ± 70 m.y. after accretion, based on Sr/Rb model ages for closure of the isotopic systems in KREEP (302). The hafnium–tungsten isotopic system (303) indicates that of the majority of the Magma Ocean solidified within 50 m.y., a time close to the lower limit of the Sr/Rb system closure uncertainty.

Cratered Highlands/Very Large Basins (Stage 3—Pre-Nectarian). The Cratered Highlands/Very Large Basin Stage of lunar evolution (Plate 7) represents the early Pre-Nectarian of the US Geological Survey time-stratigraphic system (304) and records the sustained violence of the first few hundred million years in the history of the Moon and the terrestrial planets. The stage commenced

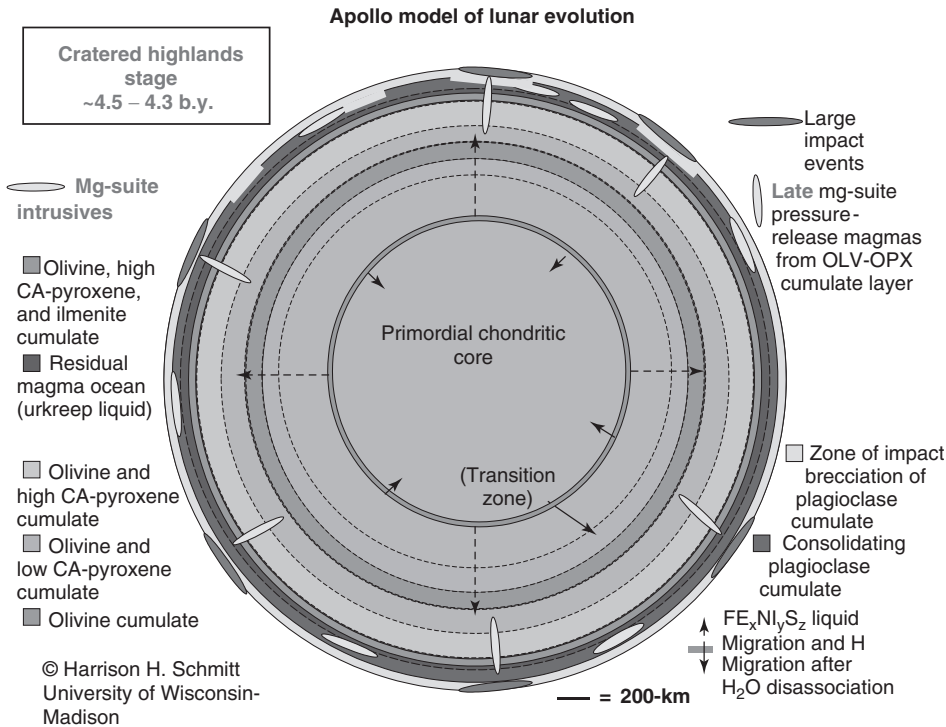


Plate 7. Apollo model of lunar evolution—Cratered Highlands Stage ~4.5–4.3 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

as the developing lunar crust gained consistent stability and coherence sufficient to retain the imprint of craters excavated largely by the continuing in-fall of residual, preaccretion debris and/or meteors, as well as comets ejected from the outer solar system. Crystallization ages of several Apollo 17 samples of the Mg-suite of rocks and Sr–Rb model ages, as discussed before, suggest that this period began about 4.5 b.y. ago and overlapped the previous Magma Ocean Stage as the last residua of the Magma Ocean continued to crystallize and differentiate. Similarly, recent Sm–Nd isochron ages for two ferroan anorthosite clasts from an Apollo 16 highland breccia are ~4.54 and 4.40 b.y. (305). A possible lower limit for the duration of the stage is 4.2 b.y. (306) at which time most K–Ar radiometric clocks in Apollo 16 breccias of apparent Pre-Nectarian age were reset. The sustained high rate of impact probably resulted from preaccretion debris remaining in crossing orbits (307,308) and the ejection of material from orbital bands in resonance with the gas-giant planets (309). There also remains a possibility for the accretion of any original moons or rings of Earth after capture of the present Moon (310) or its aggregation from a circumterrestrial disk (311). As the Cratered Highlands formed, the crust saturated with craters at a saturation size (measured by curves of crater size vs. frequency that approach a slope of -1) of 60–70-km in diameter (312) (Fig. 9). This means that a relatively uniform megaregolith, including any heterogeneities introduced into the crust by Mg-suite magmatic intrusions, developed to a depth of at least 25 km, the approximate depth to which Apollo passive seismic interpretations indicate that intense



Figure 9. Typical Cratered Highlands on the farside of the Moon (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

fracturing of the upper crust exists (313). Although average depths of about 20-km for the megaregolith have been proposed (314), brecciation to the greater depths suggested by seismic data is plausible due to the near superposition of many large craters and impact-generated fractures that extend below the transient floor of each crater. The superposition of ejecta blankets from large basins, on the other hand, will have produced much thicker regional zones of megaregolith, even though lithostatic pressure may have closed most fractures below 25 km. During this stage, regional textural homogenization of the upper crust down to the base of the megaregolith was indicated by the pre-Apollo lunar mapping program (315) and by the sample suites from Apollo 16 (316). Apollo 17's impact breccias related to the Serenitatus event (boulders at Stations 2, 6, and 7) also record crustal recycling by many earlier basin-forming and cratering events. These breccias further illustrate the processes and incomplete degree of mixing (317,318). The

regional nature of the compositional homogenization process during this period has now become broadly defined by remote sensing of the lunar highlands from the more recent Galileo (319–321), Clementine (322), and Lunar Prospector (323) spacecraft. This process, however, appears to have added less than 0.3% of an extralunar component to the Cratered Highlands (324) if iridium is considered a surrogate for that component.

The creation of the majority of large basins, discussed in the next section, followed the Cratered Highland Stage as a distinct period in lunar evolution. Two and possibly four extremely large impacts events (325–327), however, upon which later large basins are superimposed, may have occurred during this stage. Two of these formed the nearside Procellarum Basin and the farside South Pole-Aitken Basin (Fig. 10). These two basins are about 3200-km (transient crater ~ 2100 -km) and 2500-km (transient crater ~ 2000 -km [?]) in diameter, respectively. An older, third structure or cryptic basin ~ 3000 -km in diameter, centered roughly on Mare Tranquillitatis, is indicated by a partial ring of anomalously high iron plus titanium seen in the Lunar Prospector thermal neutron data between latitude $+30^\circ$ and -45° and longitude $+60^\circ$ and $+120^\circ$ (328) (Plate 5). The center of the transient crater of this very large cryptic basin, tentatively referred to as the Prospector Basin, coincides roughly with the global selenopotential low at the lunar surface. Prospector may have been responsible for the

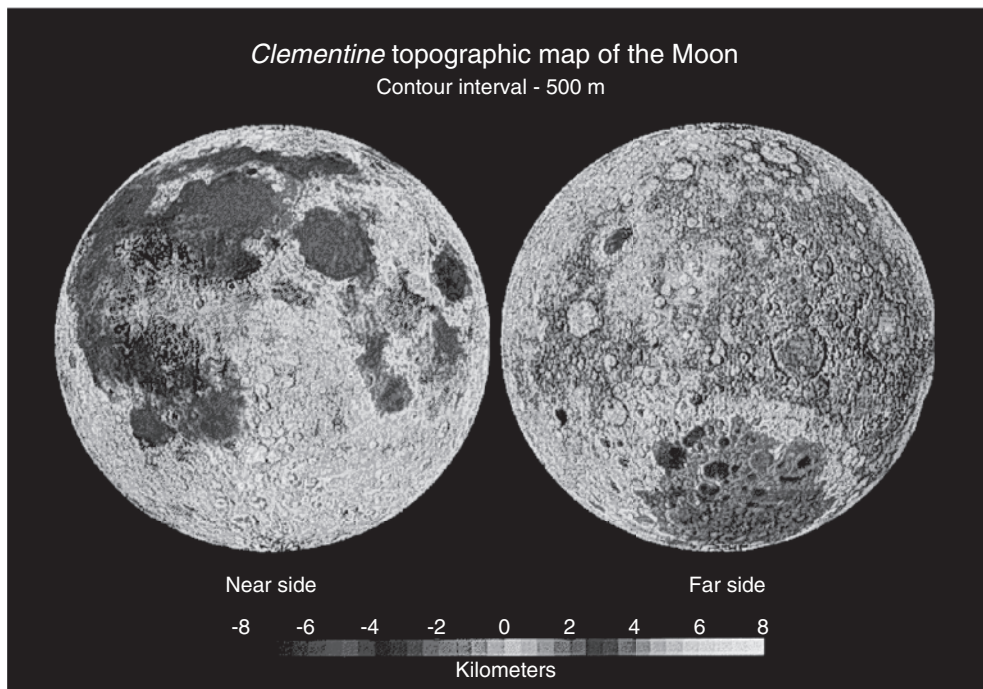


Figure 10. Topographic map of the Moon from Clementine data showing the Procellarum Basin in the upper left of the nearside image, the South Pole-Aitken Basin in the lower portion of the farside image, and the Cratered Highlands of the farside and southern nearside (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

present ~ 2 -km offset in the center of the figure from the center of mass (329) that produces this low. The lack of physiographic and strong geochemical definition of Prospector suggests that it would have been formed before full consolidation of the anorthositic crust ~ 4.5 b.y. ago and before a signature of the concentration of KREEP components in the residual Magma Ocean could be left in the crust. Without other distinguishing geochemical signatures, Prospector's ring of anomalous iron plus titanium in the crust, now nearly homogenized with the Cratered Highlands of the region, may represent a trace of partially differentiated Magma Ocean brought to the surface around the transient crater.

The Procellarum event (Plate 8) may have taken place in the midportion of the Cratered Highlands Stage, possibly about 4.3 b.y. ago, if the impact degradation of its surrounding rings is any indication. Crystallization ages near 4.3 b.y. for some of the Apollo 14 KREEP-related samples may be related to the Procellarum event. In response to such an event, KREEP-rich basalt intrusions and extrusions might have been emplaced throughout the Procellarum region (330). However, 700 m.y. later, extensive eruptions of mare basalt partially filled the Procellarum Basin and other basins contained within it (see below) and covered most materials created or emplaced as a consequence of the Procellarum event. Beneath the transient Procellarum crater, pressure-release melting of olivine and low calcium–pyroxene cumulates (harzburgite), if those cumulates were near their partial melting temperature, would have produced intrusive magmas that may have crystallized the relatively young members of the Mg-suite of samples (331). The Procellarum impact also would have removed the early formed

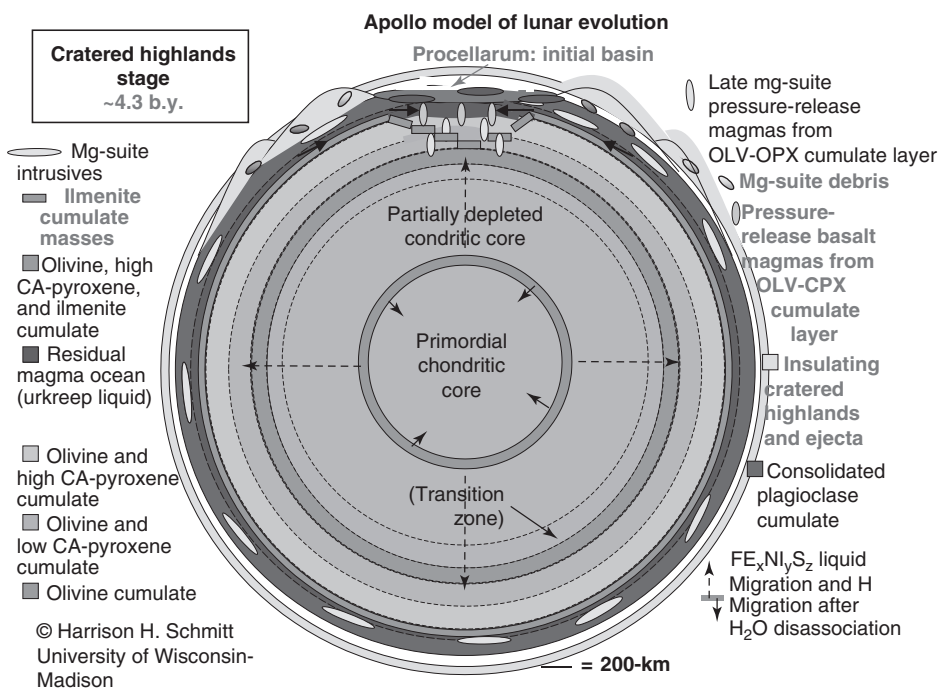


Plate 8. Apollo model of lunar evolution—Cratered Highlands Stage ~ 4.3 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Cratered Highlands from the region. Present crustal thickness modeling for the region is consistent with thinning of the crust from about 70-km to about 40-km thick (332,333). By removing a large area of insulating megaregolith and requiring that a new zone of insulating crust develop subsequently, the event also may have delayed future remelting of the underlying upper mantle to produce mare basalt magmas. Crystallization ages of mare basalts (334,335) sampled by Apollo 12 and 15 within the Procellarum Basin are younger (3.1–3.4 b.y.) than mare basalts to the east (3.6–3.9 b.y.) sampled by Apollo 11 and 17. On the other hand, the ages of mare basalts fragments in the more limited Luna 16 and 24 samples from Mare Fecunditatis and Mare Crisium, respectively, east of the Apollo 11 and 17 sites, are in the 3.3–3.6 b.y. range. An Ar–Ar age from Mare Fecunditatis basalt is 3.41 ± 0.04 b.y. (336), and two Ar–Ar ages from Crisium basalt fragments are 3.30 ± 0.04 b.y. (337) and 3.61 ± 0.12 b.y. (338). These relatively young ages may be the result of basaltic maria appearing first at the selenopotential low and then progressively away from that low (339) or of an as yet unrecognized, very large basin-forming event in that region removing older Cratered Highlands, as suggested before for Procellarum. Some suggestion for such a very large event centered on the equator southeast of Crisium is apparent in the gravity models based on Lunar Prospector data (340,341) and in the albedo maps from Clementine (342) (Clementine may be an appropriate name for this possible basin). Mare Australus consists of the basaltic lava fill of many craters and low areas in a ~1200-km diameter basin that also may indicate a very large impact.

The largely farside South Pole-Aitken Basin is the largest and oldest unambiguous impact basin on the Moon (343–348). It can be argued that the South Pole-Aitken event (Plate 9) took place near the end of the Cratered Highland Stage. For example, this event removed most of the insulating megaregolith of the Cratered Highlands and also thinned the total crust thickness to about 40 km (349); however, little new insulating megaregolith formed afterward. Gravity modeling also indicates that the lunar crust has been thickened by ejecta on the highlands surrounding South Pole-Aitken. If the Procellarum Basin is also the result of an impact, the combination of ejecta of Cratered Highlands material from South Pole-Aitken and Procellarum thickened the crust between them to about 120 km. The Galileo, Clementine, and Lunar Prospector remote sensing data have produced additional insights about South Pole-Aitken. Although actual magnesium to iron ratios have not been determined, these data suggest that most of the dark albedo surfaces in that basin, where iron, titanium, and potassium are lower than in mare basalts, may be related to impact melting of noritic or gabbroic Mg-suite rocks within the lower ferroan anorthosite crust. As in the Procellarum, Mg-suite related lavas may have been generated by pressure-release melting of the mantle's olivine–orthopyroxene cumulates beneath South Pole-Aitken. The presence of a weak KREEP signature in the South Pole-Aitken Basin may indicate that deep Mg-suite masses included some intrusives derived from Magma Ocean residual liquids of intermediate maturity. Imbrium ejecta in this region, however, cannot be eliminated as a source of KREEP-bearing material (350,351). In spite of a basin depth of 12 km below the mean lunar radius, mare basalt appears to be rare in the basin (352). South Pole-Aitken removed essentially all of the insulating surface layer without time for replenishing it, as indicated by the better definition of related physiographic

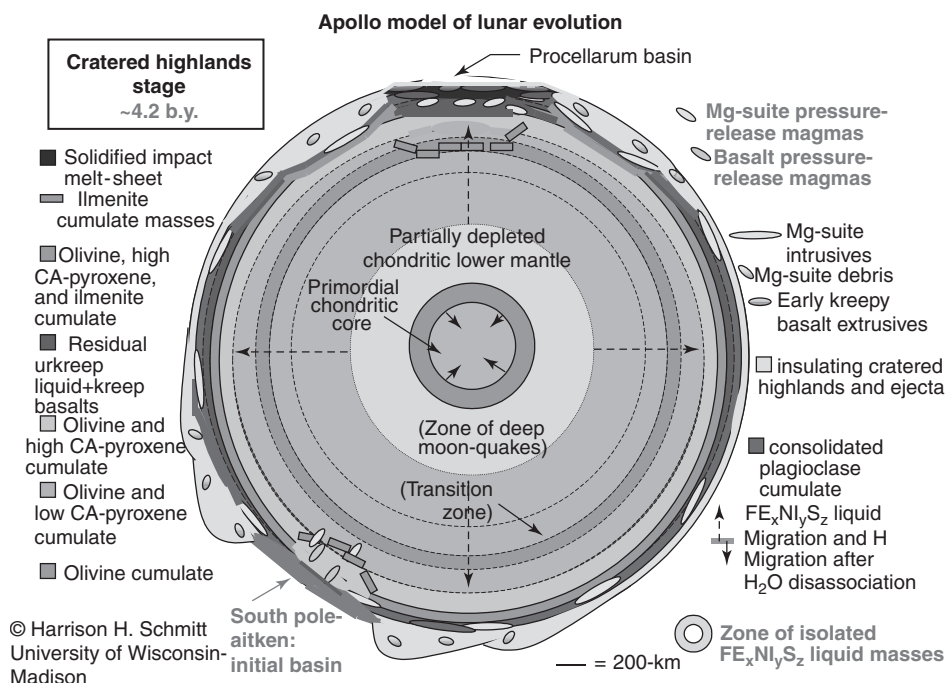


Plate 9. Apollo model of lunar evolution—Cratered Highlands Stage ~4.2 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

features relative to Procellarum. The South Pole-Aitken event thus appears to have prevented the later large-scale remelting of underlying upper mantle cumulates and the subsequent regional eruption of mare basalt, even though significant heat-producing, KREEP-related materials may be present in the underlying crust. Some local mare basalt signatures, however, have been identified in the northern portions of the basin (353), suggesting late partial melting of the underlying mantle, as occurred elsewhere on the farside of the Moon.

Thermal neutron spectra obtained by Lunar Prospector disclose a zone of high iron plus titanium, relative to other nearby crustal elemental signatures, just outside the rim of both the Procellarum and South Pole-Aitken Basins (354). These annuli suggest that material at and immediately outside of the now exposed rims of the Procellarum and South Pole-Aitken Basins came from deep crustal or upper mantle rocks. Until the magnesium content of these annuli is determined, their full significance will be uncertain; however, both events may have excavated levels of the lower crust that included abundant mafic material, possibly Mg-suite intrusives. This conclusion is consistent with new assessments of Apollo passive seismic data (355) which indicate a gradual increase in crustal density from 2.9-gm/cm³ in the upper crust to 3.4-gm/cm³ at the crust-mantle boundary. Lunar Prospector gamma-ray spectrometer data (356), on the other hand, indicate that neither of these enormous events excavated significant material related to urKREEP. This strongly suggests that urKREEP magmas had yet to move into the Moon's lower crust at the time of each event, a magmatic

episode discussed further later in connection with the Old Large Basin Substage when urKREEP apparently did move upward on a global scale.

The extensive movement of residual Magma Ocean liquids across and possibly along the crust–mantle boundary region may well have occurred beneath both Procellarum and South Pole-Aitken in response to the regional reduction in lithostatic pressure that would coincide with their formation (357–359). The coincidental formation of another large basin, the 1160-km diameter Imbrium basin near the center of Procellarum, further resulted in the redistribution of KREEP-related materials in and roughly radial to Imbrium (360). This scenario for urKREEP migration after a postulated Procellarum event provides an alternative to recent proposals of chemically asymmetrical differentiation of the Moon (361–365) to explain the present surface concentration of KREEP-related material around Imbrium. In this regard, an extreme concentration of KREEP-related material in the Procellarum region has been postulated in the form of the “Great Lunar Hot Spot” (366) now generally referred to as the “Procellarum KREEP Terrane (367). Asymmetry by the end of Magma Ocean solidification is also inherent in suggestions of the overturn of ilmenite cumulates (368) and overall mantle instabilities (369). On the other hand, if the Magma Ocean solidified in a roughly spherically symmetrical manner, such a KREEP-related concentration would not be expected as a consequence of differentiation. Like any fluid in orbital space, a Magma Ocean would tend to form a spherical shell subject to rotational, tidal, and convectional stresses. A generally spherically symmetrical solidification of this shell on the Moon is suggested strongly by (1) the pervasive evidence of both an originally global anorthositic crust, as discussed previously, and (2) globally distributed eruptions of mare basalts derived by partial melting from mantle cumulates (see Basaltic Maria discussion below). What asymmetries, if any, may have been introduced by rotational, tidal, and convectional stresses are not yet clear.

As implied before, one of the most important effects of the formation of the Cratered Highlands was a reduction in the thermal conductivity of the crust. As indicated by Apollo passive seismic data (370,371), low thermal conductivity corresponds to the high scattering and low attenuation of seismic waves that exist in this upper zone or megaregolith of the lunar crust. Thus, individual rock particles in the outer 25 km of the crust would be dominantly in point contact with each other and can be expected to be highly insulating thermally. Thermal transfer would be largely by radiation between particles rather than by conduction. The increasingly insulating character of the pulverized upper crust over that which would have prevailed during the crystallization of most of the Magma Ocean would arrest the cooling of the residual Magma Ocean during the Cratered Highlands Stage. It allowed the gradual accumulation of radiogenic heat necessary eventually to remelt partially the source regions in the upper mantle that subsequently produced the mare basalts and various pyroclastic volcanic eruptions. The resulting downward wave of heating would proceed into the upper mantle from the still molten and significantly radioisotopic urKREEP residual liquid at the base of the crust. This heating would be augmented by radioactivity retained interstitially within the deeper Magma Ocean cumulates. Proposals related to a Procellarum “hot spot” discussed before have led to a suggestion that a radiogenic heating mechanism by itself could explain the concentration of

basaltic maria on the lunar nearside (372). The thermal models developed in support of this proposal indicate that such a concentration of radiogenic heating would have kept the urKREEP zone beneath the crust molten for a few billion years. Maintaining urKREEP in a molten state only beneath Procellarum, however, does not explain the presence of isolated basaltic maria in deep farside craters (373,374). More critically, it cannot explain the persistence of gravitational anomalies due to mass concentrations (mascons) in the Procellarum Basin after 3.9 b.y. ago (see following Large Basin Stage discussion). The evidence is strong that the residual Magma Ocean had fully solidified in the crust by the time the young, mascon basins had formed. Although some significant migration of urKREEP liquids toward the Procellarum region may have occurred after the Procellarum basin-forming event, mare basalt magmas probably were prevented by the thicker farside crust from reaching the surface on the farside to the extent seen on the nearside. In the specific case of South Pole-Aitken, the loss of the insulating Cratered Highlands would have limited the remelting of the underlying upper mantle.

Large Basins (Stage 4—Pre-Nectarian–Lower Imbrium). Apollo 17 and Apollo 15 landed in the proximity of major structural features related to the formation of Serenitatis and Imbrium, respectively, two of the approximately 50 large circular basins on the Moon (Fig. 11). The extreme, but less than ~ 100 -km diameter scale cratering intensity that characterized the Cratered Highland Stage of lunar evolution had waned significantly before these and most other large basins were formed by extremely large and energetic impacts (375). This conclusion is consistent with the relatively good preservation of the concentric ring structures and some features of the surrounding ejecta blankets of most basins defined as “large,” whose diameters are greater than 300-km (376). In aggregate, crustal thickness determinations (377) indicate that these basins have redistributed large volumes of the upper 10–30-km of the fragmental debris (megaregolith) of the Cratered Highlands, particularly on the nearside of the Moon. Preservation of many of the physiographic features of the majority of large basins contrasts sharply with the vagueness or absence of similar features related to the very large basins (Procellarum, South Pole-Aitken, and possibly Prospector and Clementine) that formed during the Cratered Highlands Stage. These earlier very large basins may indicate that many more such basins formed, possibly about 14 (378), only to be destroyed by the combined effects of later basin forming and the overall, very high cratering rate during the Cratered Highlands Stage. As noted before, the Cratered Highlands are saturated with craters 60–70-km in diameter (379), a size–frequency parameter that requires a much higher rate of cratering than recorded in subsequent lunar history. Although little direct evidence of more than a few, very large “cryptic” or hidden basins yet exists and analyses of the uniformity of anorthosite distribution suggest that few actually formed (380), the previous existence of at least some cannot be ruled out, such as the possibility of Clementine, southeast of Crisium, discussed previously. One other such cryptic basin may encompass the roughly 2000-km diameter farside region that includes the smaller basins Mendeleev, Moscoviense, and Freundlich-Sharonov. This region appears to have a crustal thickness 20–30-km less than surrounding highlands (381) but no other indications of a very large basin event.



Figure 11. Near full Moon view showing examples of old and young, large basins both containing basaltic maria. The irregular basin at the lower left of center is Tranquillitatis, whose southern portion contains the Apollo 11 landing site. The circular basin at the lower left is Serenitatis. The Apollo 17 landing site is located in the lower right portion of the surrounding ring of mountains. The farside mare basin, Tsiolkovskiy, is visible just right of the top (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Large crater-forming events that formed a specific period in solar system history may have continued for several hundred million years after the less energetic background impacts of the Cratered Highlands Stage had largely ceased. A strong presumption can be made that a source of objects discrete from that which caused the earlier intense cratering was responsible for most large basins in the period between 4.2 and 3.8 billion years (382,383) or was part of a “cataclysm” at about 3.85 billion years (384,385). A discrete source is required even if a short-lived cataclysm is assumed. How long the large basin impact period lasted will be discussed further later; however, this discrete source supplied far fewer, but far more energetic objects (more massive and/or higher velocities) than the source responsible for the intense, postaccretion cratering that created the Cratered Highlands. Four possibilities for sources of the impactors of the Large Basin Stage appear plausible at this time: (1) the proto-Kuiper Belt of cometary objects whose injection into the inner solar system was the result of orbital resonance with Neptune (386); (2) the injection of Öort cloud cometary

objects at rates greater than during the last 3.9 b.y. as a consequence of perturbations by a passing stellar body (387); (3) large protomoons of Earth, swept up by the Moon during and after capture (388); and (4) Jupiter's initial interaction with the Main Belt Asteroids, or its induced breakup of the Belt's planetesimal precursor, and injection of fragments into inner solar-system-crossing orbits by orbital resonance (389). Of these possibilities, Jupiter's initial interaction with Main Belt Asteroids and/or the breakup of the original Main Belt planetesimals or ejection of Kuiper Belt objects due to interaction with Neptune appear to be the best present choices as discrete impactor sources. Modeling studies suggest that interactions between Neptune and Kuiper Belt objects probably would be completed over time frames of about 10 million years, although time frames of the order of hundreds of millions of years may be involved in the formation of the Kuiper Belt itself (390). An Öort Cloud source is currently unconstrained by modeling studies, but it seems likely that it continued to supply similar objects off and on indefinitely, given its postulated large total mass. Assimilation of protomoons of Earth, if any existed, would have involved objects of low kinetic energy relative to a coorbiting Moon, would occur over a shorttime interval, and would not explain large basin formation on other planets. At this stage of knowledge, however, it does not appear that any of these possibilities can be totally eliminated. Conversely, future studies of potential interactions of the gas giants with the Kuiper Belt and of passing stellar neighbors with the Öort Cloud might include the hypothetical constraint of large basin formation on the inner planets between about 4.3 and 3.8 billion years ago or of a short cataclysm.

Old Large Basins/Crustal Strengthening (Stage 4a—Pre-Nectarian). Globally, comparison of the old large basins with younger large basins (391) indicates that major strengthening of the lunar crust occurred during the early portion of the Large Basin Stage (392). The younger basins of the Nectarian and Lower Imbrium Systems, including Serenitatis, are circular and sharply defined structurally and physiographically. Central mass concentrations (mascons) underlie the young basins and are surrounded by mass deficiencies under mountainous rims several thousand meters high (393–395). Analysis and modeling indicate that most of the mascon signal comes from thick plates of basalt that partially fill the young basins in contrast to relatively thin basalt fill in older basins. A small part of the mascon signal may come from upward bulges in the crust–mantle boundary. A very large amount of mare basalt as hidden intrusions in the megaregolith beneath the basins may also contribute. On the other hand, the old basins are only irregularly circular, have relatively low mountainous rims, underlie ejecta from young basins, and today are largely compensated for gravitationally, that is, they have no mass concentrations or deficiencies associated with them. Thus, old large basins have adjusted isostatically, but young large basins have not (396) (Plate 10), leaving the latter with deep holes for mare fill and uncompensated for mass deficiencies beneath their ejecta rims. This suggests that the fracturing of the lunar crust by the older basin-forming events permitted urKREEP liquids to migrate into the crust. When these liquids moved upward and solidified, the potential was lost for rapid, postimpact isostatic adjustment by urKREEP magma movement at the crust–mantle boundary. A strengthening boxwork of solidified KREEP-rich intrusions also would be created

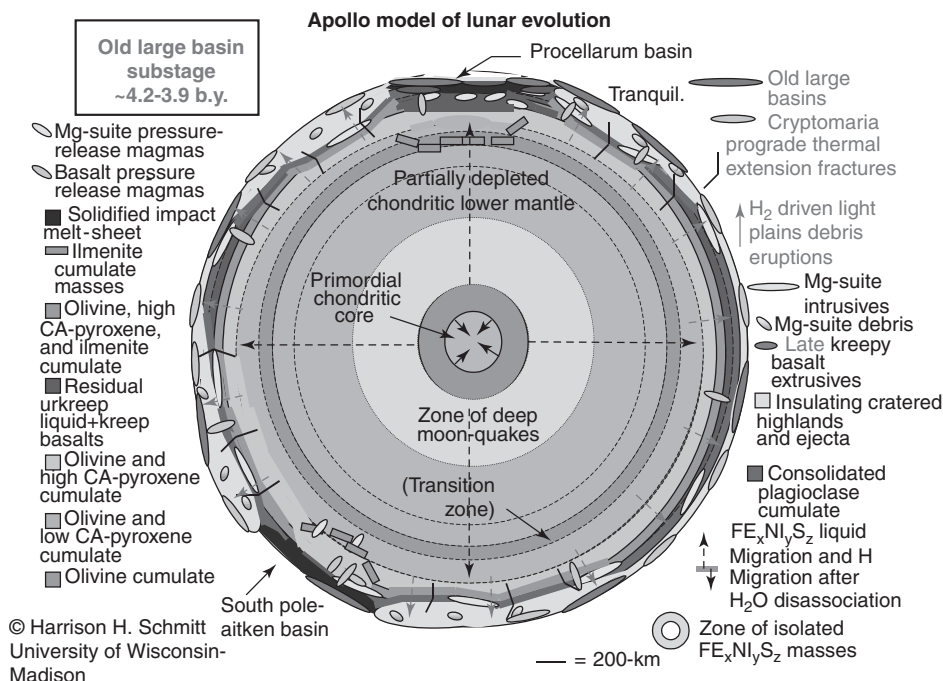


Plate 10. Apollo model of lunar evolution—Old Large Basin Substage ~4.2–3.9 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

in the lower crust. Although old basins also have some mare fill, higher basin floors would result in thinner mare basalt plates than in younger, unadjusted basins. If this hypothesis is correct, the remnants of the Magma Ocean remained in liquid form until about 3.92 ± 0.03 b.y. ago. This is the apparent age of Nectaris (397,398), apparently the oldest basin that has an uncompensated for mascon (399). The Apollo sample suite lacks clear-cut examples of ferroan anorthosite in association with material that might be interpreted as urKREEP intrusive. Except for Apollo 15 KREEP basalts, rocks that have a strong KREEP chemical component have been found only as impact melt breccias. These two observations (400,401) suggest that urKREEP intrusives are confined to the lower crust and have been exposed only in a single deep excavation, namely, Imbrium, where impact melting predominated. Probably because of the thinning of the upper crust by coincidence of the Imbrium transient crater with that of the Procellarum event, 3.83–3.86 b.y. old mare basalts (402) mixed with older KREEP-rich material may have reached the surface in that region roughly coincident with the Imbrium event at 3.82 b.y. (403).

Recent studies (404) have increasingly extended the areas of the Moon in which craters that penetrated young large basin eject deposits have exposed older dark materials that have basaltic characteristics, the so-called “cryptomaria” (405–408). The cryptomaria are related in age to the old, large basins because they also lie under ejecta from the young, large basins which is why they are “cryptic.” The volume of these apparently basaltic lavas of pre-Young Large Basin age is uncertain, but may be of the order of 10^6 -km³ (409) compared with basaltic

maria volume estimates of the order of 10^7 -km³ (410). Cryptomaria may be represented in the Apollo samples by basalts of ages clearly greater than 3.92 b.y. (411,412) or by KREEP-related basalts whose model ages are 4.2–4.4 b.y. (413). Most pre-3.92 b.y. basalts have high aluminum content, and others have KREEP affiliations and a variety of unusual chemical combinations (414). This chemistry may be explained by derivation from either urKREEP residual liquids or pressure-release partial melts of shallow Magma Ocean cumulates positioned between 100-km and 200-km depth and just below the radioisotope-rich, residual liquids of the Magma Ocean. Movement, intrusion, and eruption of magmas derived from residual liquids at the crust–mantle boundary and deeper pressure-release melting of a mantle near its partial melting point would occur in response to structural disruptions and the removal of tens of kilometers of overlying Cratered Highlands megaregolith. Such disruptions would accompany extraordinary events like those that created Procellarum, South Pole-Aitken, and other old, large basins. Future comparisons of the details of Lunar Prospector's gamma-ray spectrometer data and of relatively large areas of identified cryptomaria exposure may help further discriminate and define their source or sources, particularly if a KREEP signature becomes apparent. Such a signature would be expected for magmas from or passing through zones near the base of the crust. If it turns out that the compositions of cryptomaria are not significantly related to KREEP, it would suggest very rapid ascent of their magmas or their eruption in the narrow time window before the formation of young large basins and after the solidification of urKREEP liquids in the lower crust. The identification of cryptomaria on a global scale has other implications. The eruptions of early partial melts or urKREEP residual liquids from beneath the lunar highlands may have included significant residual volatiles and been at least partially pyroclastic. As a consequence, they may have contributed to the formation of light plains deposits or Cayley Formation units (415,416). Such smooth, light colored units in the lunar highlands actually may have several origins. In nearside regions affected most by young large basin events, they may be largely the result of the uniform settling of material in fluidized debris flows that extended beyond zones of continuous ejecta. On the other hand, those deposits observed by the author in old craters in the farside highlands may be the result of internal processes. An origin by pyroclastic eruptions dominated by entrained, fine-grained crustal debris is suggested by the smoothness of these plains, their similarity in albedo to other highland surfaces, irregular rimless depressions in their surfaces, and the absence of nearby young, large basins as sources of fluidized debris.

The formation of each of at least 29 old, large basins, including Procellarum and South Pole-Aitken, had five, essentially global effects. First, upper crustal material was redistributed regionally as ejecta blankets, debris flows, and impact melt flows and intrusions. Second, excavation and broad distribution of Mg-suite rocks and impact melts from solidified intrusions in the lower crust occurred. Third, regional thinning and thickening of the crust took place. Fourth, the lower crust beneath each basin was fractured extensively. And fifth, local areas of temporarily low lithostatic pressure were created under each basin, and mantle material at or near its partial melting point was commensurately mobilized. After each large basin formed, the relatively low-density, immediately underlying, urKREEP liquids would tend to move upward into the fractured lower crust

and brecciated upper crust. In the process, those liquids, probably with small amounts of superheat due to pressure release, would assimilate and mix with both ferroan anorthosite and differentiated Mg-suite masses, creating the diversity of parent magmas indicated by KREEP-rich samples from the Apollo landing sites (417). As they cooled, significantly contaminated KREEP-related magmas crystallized into networks of compositionally varied intrusions. In the Procellarum and possibly other deep old basins, KREEP-related magmas and pressure-release magmas may have reached the floor of the basins only to be largely buried by younger, large basin ejecta and mare basalt eruptions. Pre-Nectaris crystallization ages of 4.2 b.y. (418) for some olivine-rich and aluminous KREEP-related basalts are consistent with this scenario.

Apollo 17 sampled some of these effects of the Old Large Basin Stage during the examination of boulders at the base of the Massifs of the Valley of Taurus-Littrow. The Serenitatis basin-forming impact partially exposed a complex sequence of roughly layered ejecta blankets (Table 2, Plate 11) in the massif walls. Layers probably include significant ejecta from at least three, old Pre-Nectarian basin-forming events, Procellarum, Fecunditatis, and Tranquillitatis, as well as the younger basins, Serenitatis and Imbrium. It is possible that samples from these ejecta units are included in those collected at Stations 2, 3, 6, and 7 at the bases of the South and North Massifs. The most likely possibilities are samples from Boulder 1 at Station 2 (419). Rb–Sr isochron ages between 4.2 and 4.0 b.y. for some of the clasts in this boulder suggest that the probably older Procellarum ejecta is not represented unless that event was the source of the fine-grained plagioclase clasts whose Rb–Sr isochron ages are about 4.4 b.y. (420).

Young Large Basins (Stage 4b—Nectarian–Lower Imbrium). Between 3.92 ± 0.03 b.y. (Nectaris event) and 3.80 ± 0.05 b.y. (Orientale event), 14 large basins formed (421,422) in a lunar crust that could support significant mass concentrations and deficiencies indefinitely (Plate 12). Alternatively, it has been suggested (423–425) that a lunar cataclysm between 3.9 and 3.8 b.y. created these basins and also the 29 or more older, nonmascon basins and the Cratered Highlands as well. This latter hypothesis rests on the predominance of ages in that range that have been measured for impact glasses in Apollo, Luna, and lunar meteorite samples from the highlands and from basin-related materials. In this context, note that all Apollo and Luna sampling took place within the regions most affected by young, large, basin-forming events and a predominance of ages related to those events would be expected. This influence also may have extended globally to affect the lunar meteorite samples. Advocates for a cataclysm argue that the influence of impact melting and thus the resetting of radiometric ages is limited to the near vicinity of the basins. This may not be the case as lunar mapping (426) and experimental results (427) suggest that very large amounts of impact melt form within the transient crater of a basin, and also such melt is widely distributed on and in ejecta. Ejected melt particles of the highest energies will have a global reach in large basins. Even so, impact melts from the highlands sampled by Apollo 16 show formation ages up to 4.2 b.y. (428), and, as noted before, Rb–Sr ages for clasts in breccias sampled by Apollo 17 also show ages up to 4.2 b.y. The question whether 50 or more large basins formed in about 100 million years or less or in about 400 million years will ultimately be determined by more extensive sampling of the lunar highlands.

Table 2. Local Sequence from Younger to Older of Pre-Mare Basalt Stage Events in the Vicinity of the Valley of Taurus-Littrow (For Regolith Depth: Shallow = 0.01–1 Meters, Moderate = 1–10 Meters, and Deep = 10–100 Meters)

Depth or thickness	Characteristics	Age or duration, b.y., Δ = time interval
~ 10-m	<i>In situ</i> brecciation and gardening from primary and secondary impacts.	$\Delta 3.87$ b.y.
~ 100-m (621)	Overtured, late-stage Cratered Highlands ejecta and/or debris flow deposits from the Imbrium event, consisting of ferroan anorthosite fragmental breccia with KREEP-related and Mg-suite clasts (Boulder 1 at Station 2) (622). Seismic velocities of ~4000-m/s (623) for the material below the basalt fill of the valley indicate relatively unconsolidated debris.	3.85 ± 0.02 (624)
Shallow	<i>In situ</i> brecciation and gardening from primary and secondary impacts.	$\Delta \sim 0.050$ b.y.
< 100-m ^a	Crystallized melt breccia from the Serenitatis event both as an extrusive sheet and as intrusives (Boulder 2 at Station 6) (625).	3.87 ± 0.08 (626)
~ 700-m ^a (627)	Overtured, late stage Cratered Highlands ejecta and/or debris flow deposits from the Serenitatis event, consisting of ferroan anorthosite fragmental breccia that contains lower crustal Mg-suite materials (Boulder 1 at Station 6) (628).	3.87 ± 0.08 (629)
Moderate	<i>In situ</i> brecciation and gardening from primary and secondary impacts.	$< \Delta 0.1$ b.y.
< 100-m ^a	Debris flow deposits of late Cratered Highlands ejecta from the Crisium event, consisting of ferroan anorthosite fragmental breccia that possibly contains KREEP and Mg-suite materials.	3.9–4.0 b.y.
Deep	<i>In situ</i> brecciation and gardening from primary and pre-Crisium secondary impacts, particularly secondaries from Smythii and Nectaris.	$< \Delta 0.2$ b.y.
< 100-m ^b	Overtured, late stage Cratered Highlands ejecta and Procellarum ejecta units from the Fecunditatis event, consisting of ferroan anorthosite fragmental breccia that possibly contains lower crustal materials including the Mg-suite (Boulder 1 at Station 2?).	(?)
Shallow	<i>In situ</i> brecciation and gardening from primary and pre-Fecunditatis secondary impacts.	$< \Delta 0.1$ b.y.

Table 2. (Continued)

Depth or thickness	Characteristics	Age or duration, b.y., Δ = time interval
$\sim 700\text{-m}^b$ (630)	Overturned, late stage Cratered Highlands ejecta and Procellarum ejecta from the Tranquillitatis event, consisting of ferroan anorthosite fragmental breccia that possibly contains lower crustal materials including the Mg-suite (Boulder 1 at Station 2?).	3.94 ± 0.06 b.y. (631,632)
$\sim 10\text{--}15\text{-km}$	<i>In situ</i> brecciation and gardening from Pre-Nectarian primary impacts and Nectaris and other pre-Tranquillitatis secondary impacts.	$< \Delta 0.3$ b.y.
10–20-km	Overturned, early stage Cratered Highlands ejecta and/or debris flow deposits from the Procellarum event [$\sim 2100\text{-km}$ in diameter]. At Taurus-Littrow $\sim 400\text{-km}$ from crater rim (633), ejecta consisted of ferroan anorthosite fragmental breccia possibly containing lower crustal Mg-suite materials.	~ 4.3 b.y.
$\sim 10\text{--}15\text{-km}$	Early Cratered Highlands ferroan anorthosite fragmental breccia.	4.4–4.3 b.y. (634)

^aRelative age relationships between Crisium and Serenitatis not established.

^bRelative age relationships between Tranquillitatis and Fecunditatis not established.

The vast majority of the nonmare materials accessible to Apollo 17 in the Valley of Taurus-Littrow represents large basin ejecta derived from the Cratered Highlands or the upper lunar crust. Using the relative ages of the large basins (429), approximate ejecta thickness expected from such basins (430), and radiometric ages determined for specific impact-related events near Taurus-Littrow, the general characteristics of the local sequence of events are given in Table 2 and Plate. 11. Investigations of the large boulders of impact melt breccia at Station 6 and 7 at the base of the North Massif in the Taurus-Littrow area documented the complexity of processes associated with large basin-forming events. The prevalence of melt breccia in the boulders suggests that they tie genetically to the 740-km diameter Serenitatis basin whose rim is 15–20-km to the west. The track to the largest of the boulders (Fig. 12) indicated that it had rolled down from a break in the slope $\sim 500\text{-m}$ above the valley floor, $\sim 1000\text{-m}$ below the top of the North Massif (431), and probably near the base of Serenitatis ejecta. The higher, $\sim 25^\circ$ slopes include outcrops or near-outcrops; some appear roughly horizontal in aspect. For the first time on the Moon, this boulder permitted observations and sampling across a major lithologic and structural contact related to a large impact (Fig. 13). A finely crystalline melt breccia—tan-gray and highly vesicular (Fig. 14)—intrudes, is in, or was laid down in contact with crystalline nonmelt breccia—blue-gray, nonvesicular and clast-rich (Fig. 15). An approximately 1-m wide

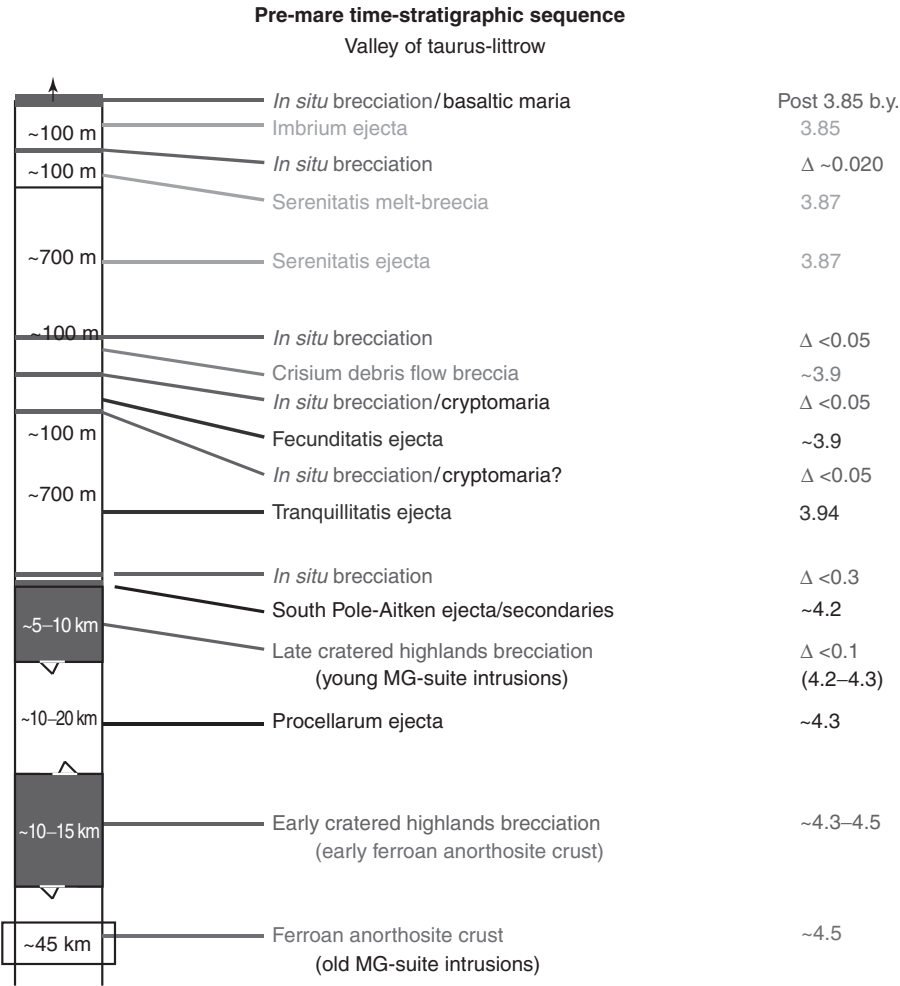


Plate 11. Premare time-stratigraphic sequence—Valley of Taurus-Littrow. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

zone in the nonmelt breccia contains small vesicles at and near the contact with the melt breccia. At Station 7, a vein of crystalline melt breccia cuts across the contact zone and through a clast, suggesting that the melt breccia may be largely intrusive. Laboratory investigations of the samples from these boulders indicate an age of crystallization for the melt breccia of 3.87 ± 0.08 b.y. (432), giving the probable age of the Serenitatis basin-forming event. The petrographic and chemical characteristics of the melt and nonmelt breccias indicate that each is polymict (multiple types of fragments) and have crystalline matrices. The matrices show a continuum in texture between very finely crystalline to poikilitic (433), reflecting the extreme brecciation, heating, melting, and mixing effects of large basin formation (434). No materials, related to the post-Serenitatis, Imbrium basin-forming event, were obvious at the base of the North Massif, although other crystalline breccias similar to the two types discussed before are present. Elsewhere, a

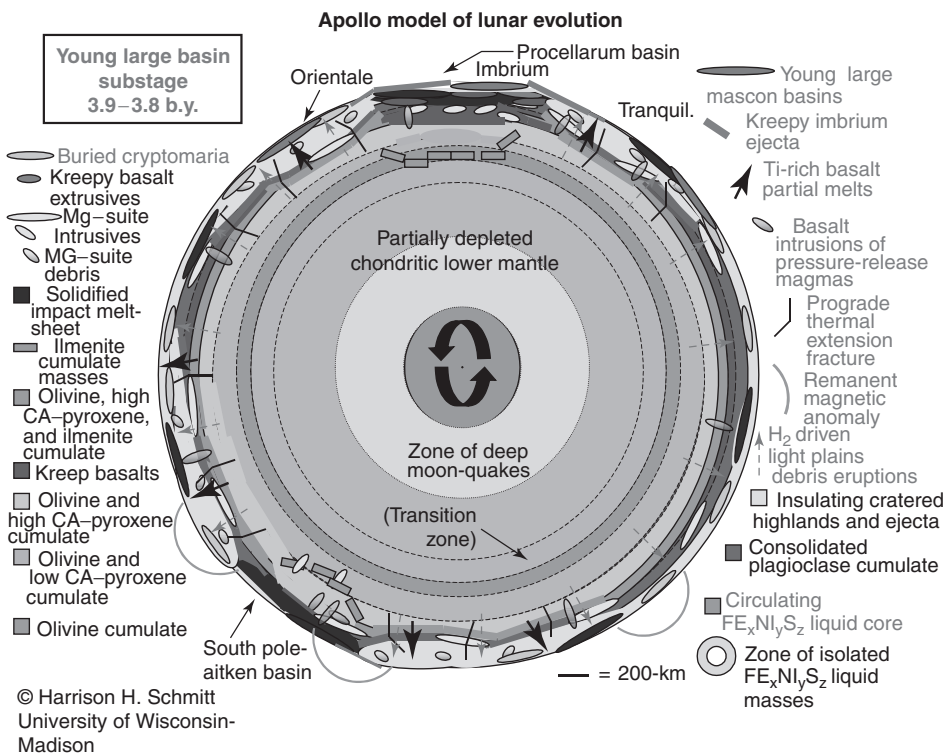


Plate 12. Apollo model of lunar evolution—Young Large Basin Substage 3.9–3.8 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

boulder of layered, relatively unmetamorphosed and less coherent blue-gray breccia investigated at the base of the South Massif appeared to have been derived from sources in the blue-gray unit at the top of this mountain. The apparent several hundred meter thickness of that unit (435,436) is consistent with the ejecta thickness estimated for points at this distance from the Imbrium Basin (437). Further, the presence of a KREEP component in the boulder's chemistry (438) suggests an Imbrium source area because most KREEP-related materials at the lunar surface appear to be associated with ejecta from that basin. The study of Boulder 1 at Station 2 suggests that the 3.85 ± 0.02 b.y. old Imbrium event (439) deposited material from the ~ 4.3 b.y. old Procellarum basin as well as recycled ejecta derived from the old large basin, Tranquillitatis, to the south, dated at 3.94 ± 0.06 b.y. (440,441).

One of the questions raised about the crystalline melt breccias (442) investigated at Stations 6 and 7 in the Valley of Taurus-Littrow relates to the vesicles or smooth-walled holes they contain. The holes formed when the breccia was partially molten and an immiscible fluid phase separated and coalesced to shape them. Elongation of the vesicles indicates that the melt breccia was still flowing when this phase was present but not at temperatures sufficient to melt all silicate clasts. Vesicles also formed in the more solid rock in contact with the melt breccia. Vesicles are common phenomena in terrestrial lavas where a variety of



Figure 12. The large boulders investigated at Station 6 at the base of the North Massif. The five boulders accessible for examination and sampling are apparently all part of one boulder that broke apart as it came to rest after rolling ~ 1.5 km down the side of the valley (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

volatile components, particularly water and carbon dioxide, are available to form the immiscible fluid. The chemical nature of the vesicle phase in lunar breccias, however, is constrained (1) by the lack of evidence that water or carbon dioxide ever existed in these lunar rocks, (2) by the absence of any discernible alteration of the minerals lining the vesicles, and (3) by the absence of anomalous mineral precipitates on the vesicle walls. A number of lines of reasoning suggest that the apparently inert fluid in question consisted largely of hydrogen and/or carbon monoxide; some helium and volatile nitrogen and other carbon compounds were possibly included. Throughout the formation of the Cratered Highlands, much of the pulverized ferroan anorthosite would have been as fine as the present-day surface regolith due to oversaturation with impacts smaller than those that formed the saturation size, 60–70 km diameter craters. This fine regolith would have been exposed to the solar wind of that time. Today, the solar wind is 96% protons (443) and may have been much more intense during the early history of the solar system, although the particularly intense T-tauri phase of solar evolution probably preceded the Cratered Highlands Stage (~ 4.4 – 4.2 b.y. ago) of lunar history (444,445). Plagioclase, the mineral phase comprising about 85% of the lunar highlands, is a mineral known to be capable of incorporating hydrogen ions in its crystal lattice (446). Epithermal neutron spectra from Lunar

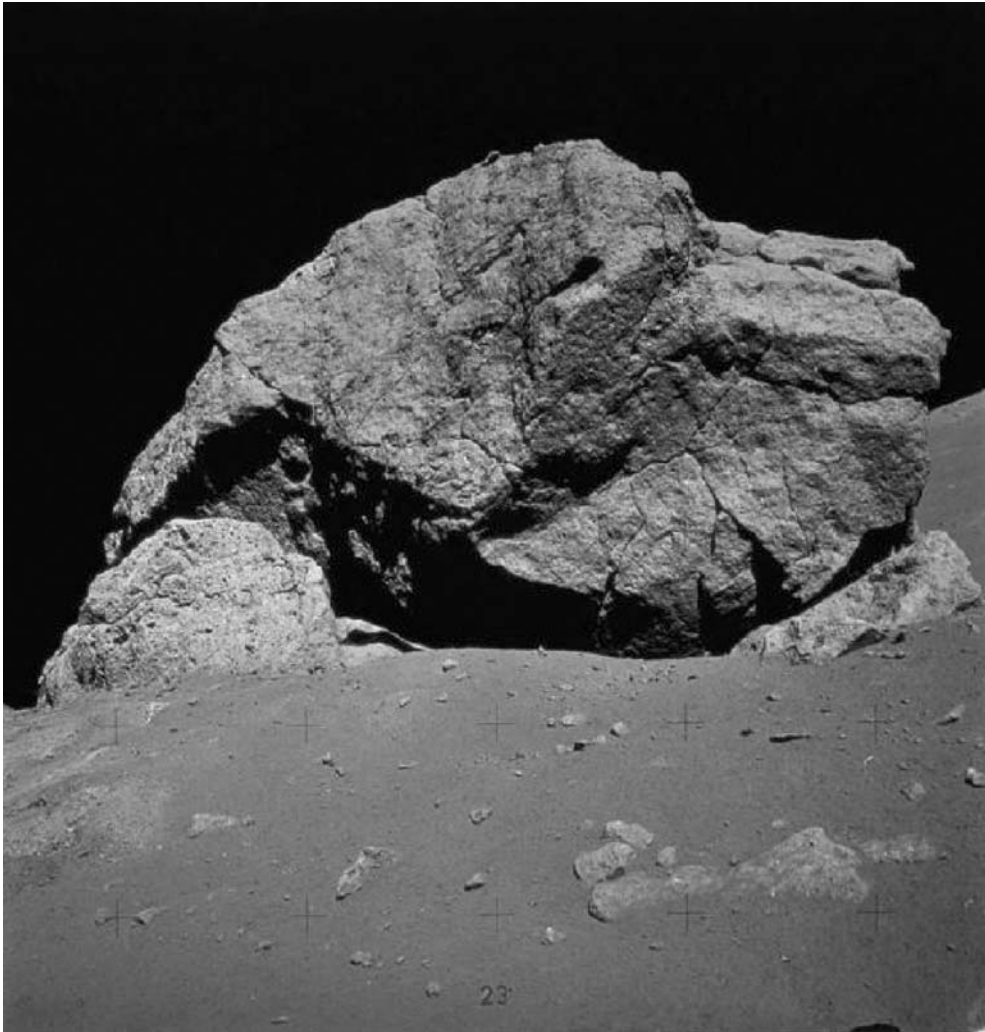


Figure 13. Contact between tan-gray, vesicular impact melt breccia (on the left) and blue-gray, clast-rich, nonvesicular impact breccia (on the right) in the boulder at Station 6 shown in Fig. 12 (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Prospector (447) indicate an average of about 30 ppm hydrogen in the present-day highland regolith (448). Thus, there is the potential for retaining significant hydrogen, as well as impactor-derived volatiles, throughout the fine portions of the crustal megaregolith. Until it is possible to sample the fines of the deep megaregolith directly, it will remain uncertain whether there was sufficient solar wind exposure during the Cratered Highland Stage to provide the volatiles necessary to create the observed vesicles. A possible alternative source of vesicle-forming fluids may be the volatiles derived from impactors or material volatilized upon impact; however, most evidence indicates that such extremely high temperature material is lost to space (449–451) or would be expected to condense

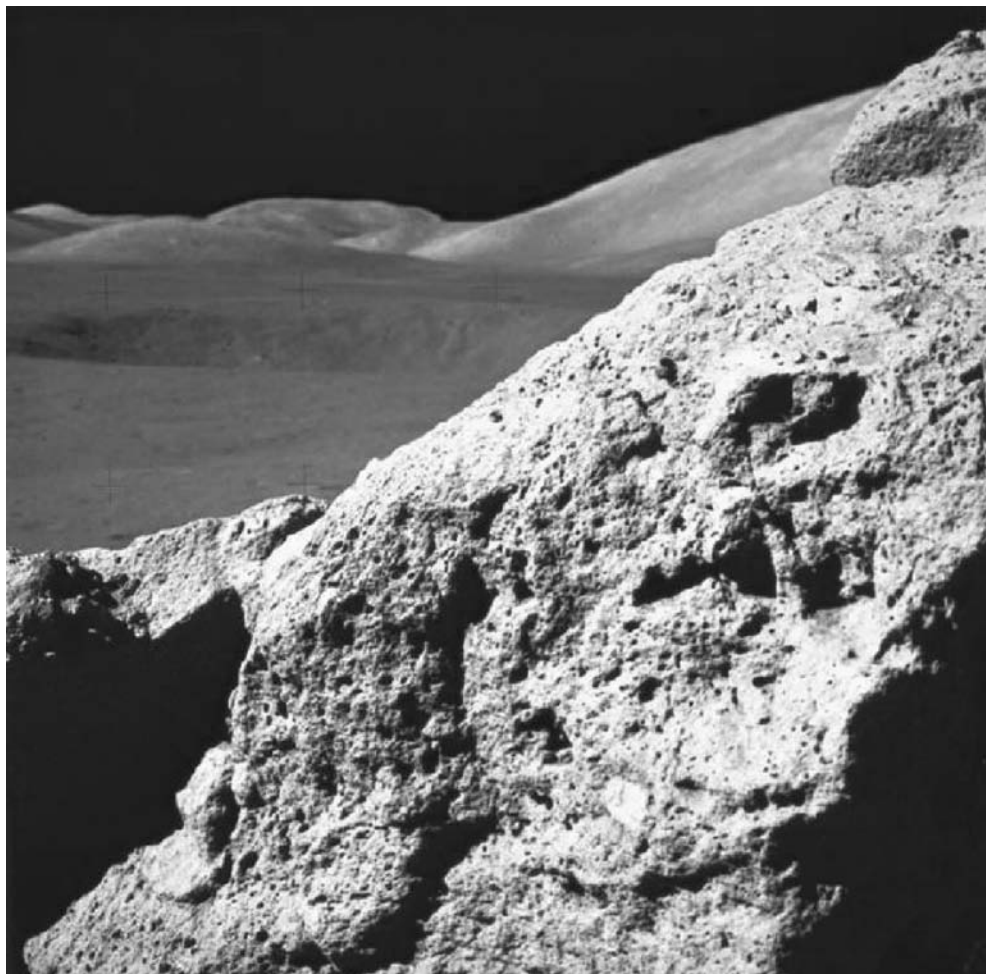


Figure 14. Tan-gray, vesicular impact melt breccia of the boulder at Station 6 shown in Fig. 12 (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

very quickly into solid material. Thus, solar wind hydrogen and lesser amounts of solar wind helium and carbon monoxide from the fine fraction of the megaregolith could be incorporated into impact melts and form the observed vesicles. Such hydrogen also may be mobilized in the crater ejecta, lowering the overall viscosity and contributing to the mobility of debris flows (452).

An additional effect of young, large basin formation appears to have been the creation of bright swirls (Fig. 16) antipodal to at least the four youngest basins, Crisium, Serenitatis, Imbrium, and Orientale (453). Another, less precise correlation, may be the large bright swirl, Reiner Gamma, which is roughly antipodal (approximately 50°W, 20°N) to the young basin, Tsiolkovskiy. The extensive region of swirls east of the Crisium basin and antipodal to Orientale were observed and photographed in detail during the Apollo 17 orbital mission. Bright swirls have no apparent topographic relief or texture, are

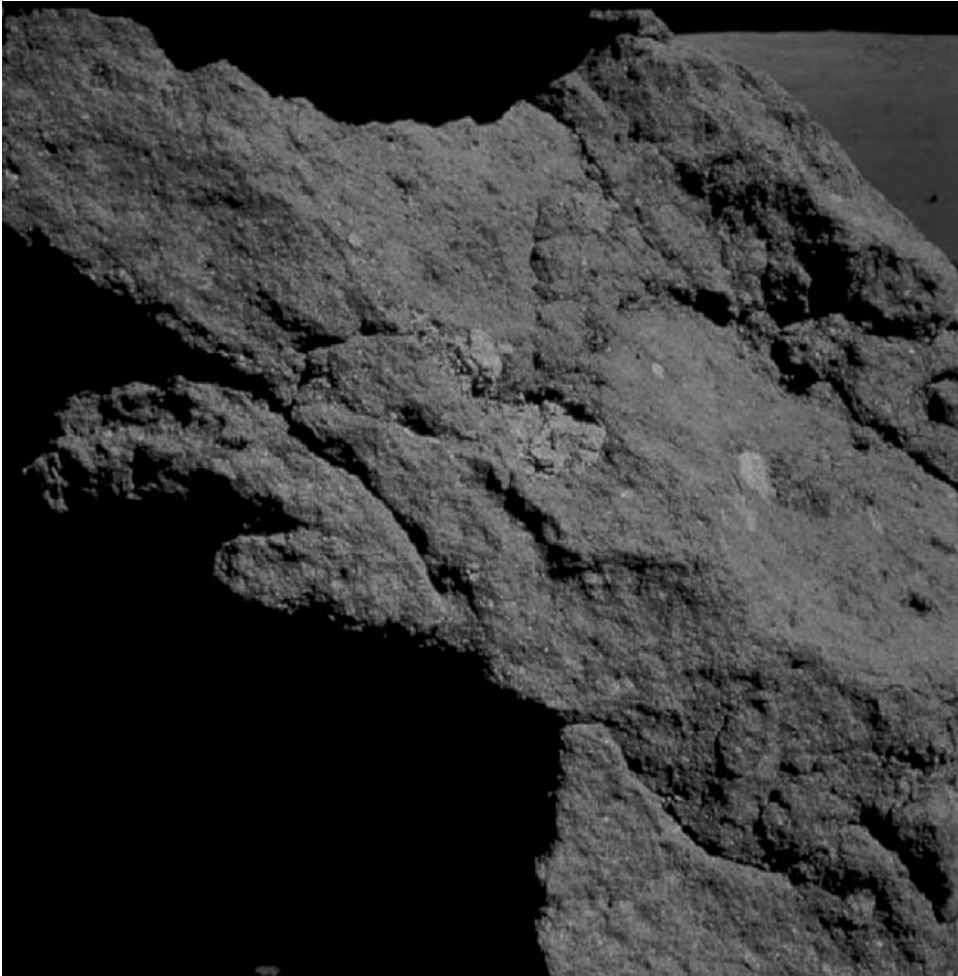


Figure 15. Blue-gray, clast-rich impact breccia of the boulder at Station 6 shown in Fig. 13 (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

diffusely bordered, and have local dark cores. There have been recent suggestions that bright swirls represent variations in proton-induced darkening of surface materials. These variations might be caused by deflections of solar wind associated with remnant magnetic anomalies that are also roughly antipodal to at least four and possibly six of the young basins (454–456). Alternatively, the swirls may be the result of an internally originated process that causes changes in the relative abundance of fine and coarse particles in the regolith (more coarse particles give a brighter albedo). This change may possibly be related to gas flow from the crust or deeper interior of the Moon that moves fine particles upward to the surface in response to seismic shaking at antipodes from large impacts. Such an origin would account for the reported lower maturity index for the Reiner Gamma swirl relative to the host mare basalt (457), although magnetization of thick, underlying Imbrium ejecta has

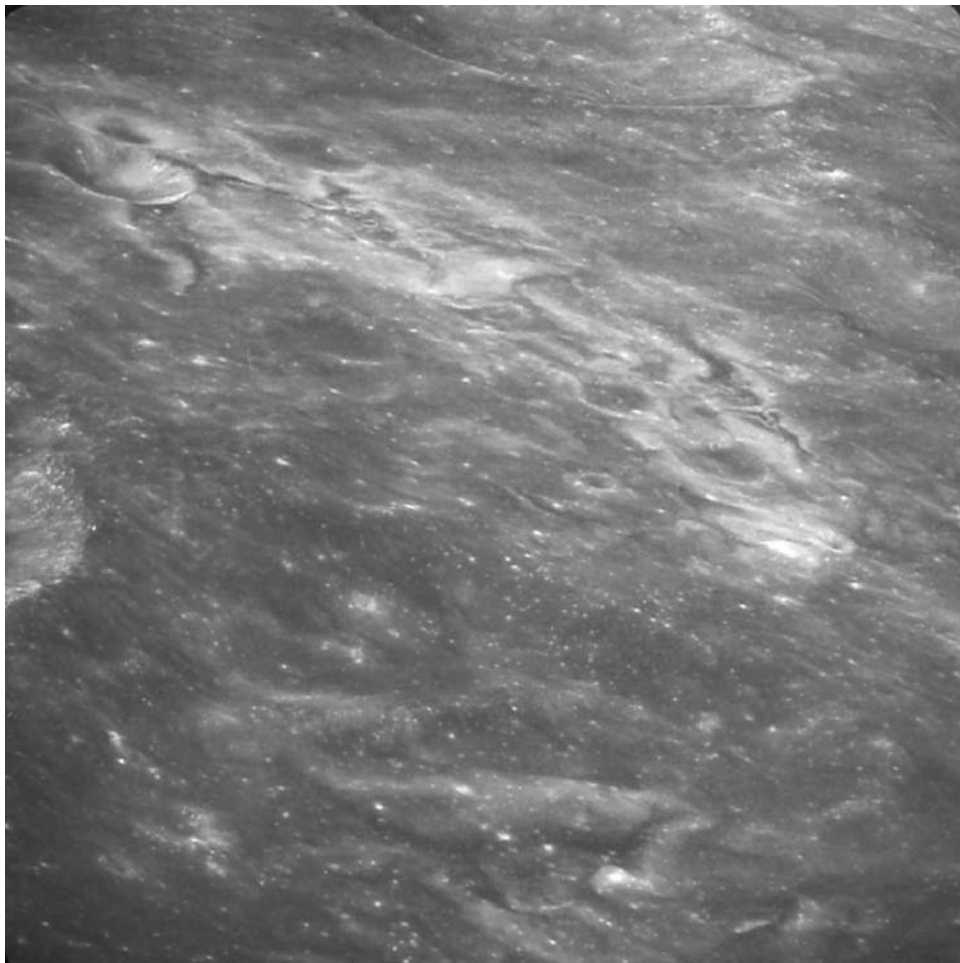


Figure 16. Light-colored swirls on the surface of the farside of the Moon, east of the Smythii Basin and antipodal to the Orientale Basin (courtesy of NASA).

also been suggested as the basis of this and two other swirls in the region (458). On the other hand, the antipodal association of magnetic anomalies, if genetically related to basin formation, would provide a constraint on the duration of a past lunar dipole field. As anomalies do not appear to be antipodal to the old large basins and are of lower intensity antipodal to Orientale, then the dipole may have been active only between about 3.92 and 3.80 b.y., the span of apparent ages of these young basins. If this suggestion survives further scrutiny, these dates may indicate the approximate times of lunar core formation and termination of core circulation, respectively. If it took about 600 m.y. for the $\text{Fe}_x\text{Ni}_y\text{S}_z$ liquid that formed the lunar core to migrate through the lower mantle and coalesce as a core, then this would be further support for the hypothesis that the whole Moon never melted during the Magma Ocean Stage. Otherwise, $\text{Fe}_x\text{Ni}_y\text{S}_z$ liquid would have formed a core nearly simultaneously with melting of the whole Moon.

Basaltic Maria (Stage 5—Upper Imbrium). Considerable evidence exists of eruptions of basaltic lavas on the Moon (Plate 13) as early as 4.2 b.y, the cryptomaria (459), and as late as 1.3 b.y. in Oceanus Procellarum (460). The majority of the partial filling of many of the large basins by mare basalt (see Fig. 11), however, particularly on the nearside of the Moon, appears to have begun about 3.8 b.y. ago and lasted for about 800 m.y. years. This period constitutes the bulk of the Basaltic Maria Stage of lunar evolution (461–463) (Fig. 17). Recent estimates of the volume of mare basalt range up to 10^7-km^3 (464), and the mean volume per eruption was about 200-km^3 (465). These volume estimates do not include potentially abundant intrusions in the pervasively fractured lunar crust. Hidden intrusives in the thick crust may account for the “missing” mare basalt on the lunar farside. The fundamental characteristics of the samples of mare basalts collected by the Apollo astronauts are their chemical and mineralogical diversity (466–469). The mare basalt sample suites from the Apollo 12 and Apollo 17 sites, however, contain many samples that appear to be largely derived from one or two individual flows or cooling units (470–472) and are discussed in more detail later. The variety of mare basalt compositions attests to the comparable variety of depths, mineral assemblages, and degrees of partial melting involved in the lunar mantle, as the mare basalt magmas formed (473,474). This variety probably was further enhanced by the differing degrees of internal differentiation that resulted from fractional crystallization and devolatilization during ascent and

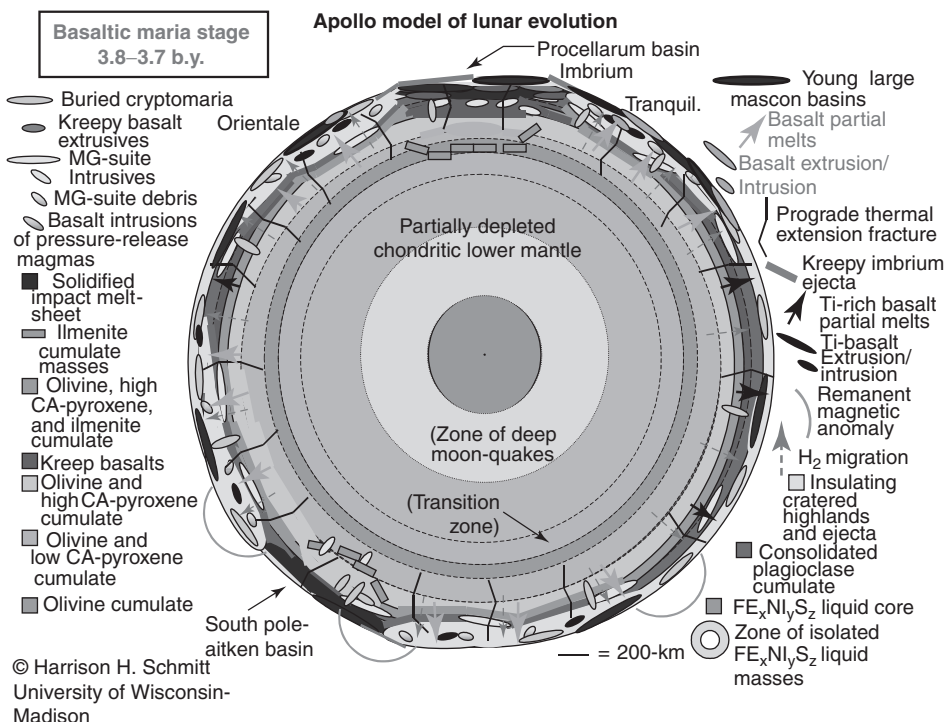


Plate 13. Apollo model of lunar evolution—Basaltic Maria Stage 3.8–3.7 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

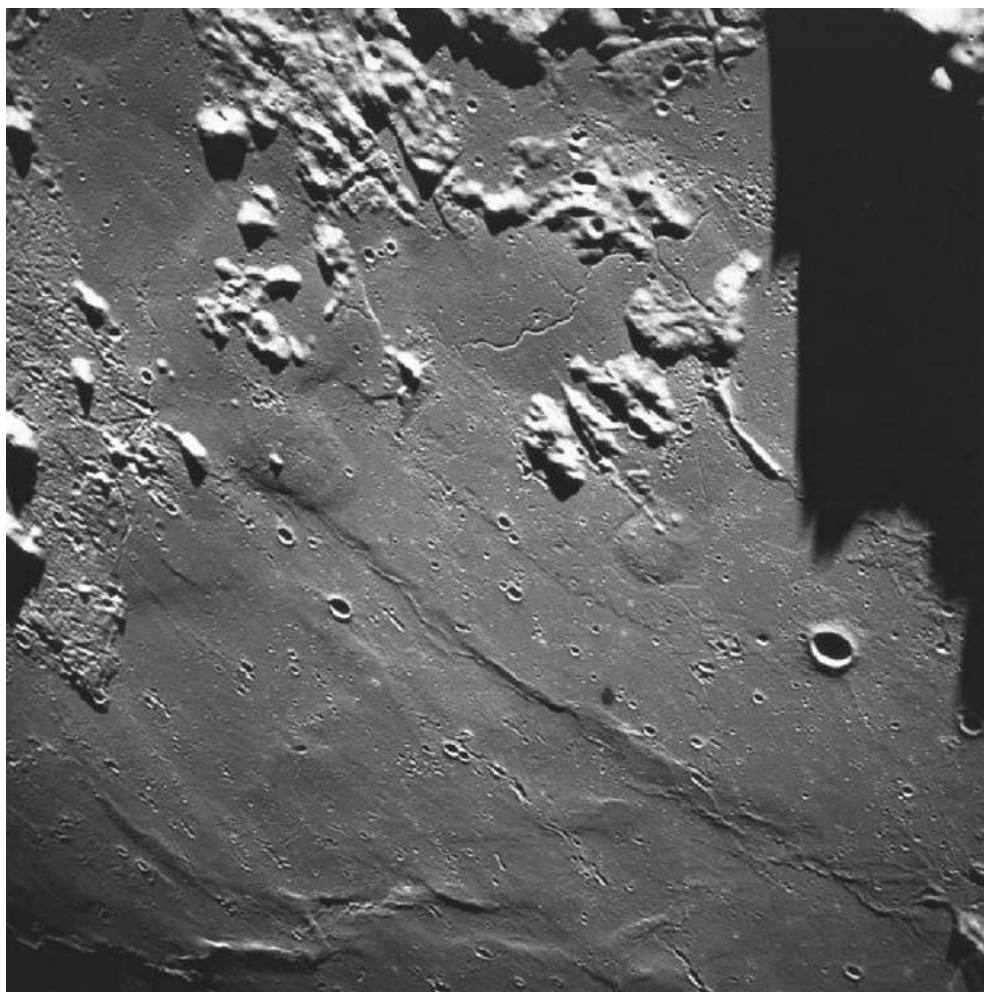


Figure 17. Mare structures near sunrise on the mare basalt surface of the eastern Imbrium Basin (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

solidification which it has been noted, also took place in the Apollo 12 basalt flows (475) (Plates 14–16).

The insulating nature of the Cratered Highlands and the heat sources in KREEP-related intrusives within and beneath the crust would have resulted in a downward wave of partial melting, progressing through the Magma Ocean cumulates of the upper mantle. This constitutes the simplest and currently most plausible scenario for the generation of mare basalt magmas during their peak eruptive period. Additional sources of heat would have included interstitial radioisotopes within the cumulates. Partial melting at increasing depths would partially reverse the sequence of cumulate formation in the cooling Magma Ocean, initially producing primary magmas rich in titanium and KREEP-related components at about a 200-km depth, followed by melting at increasing depths to

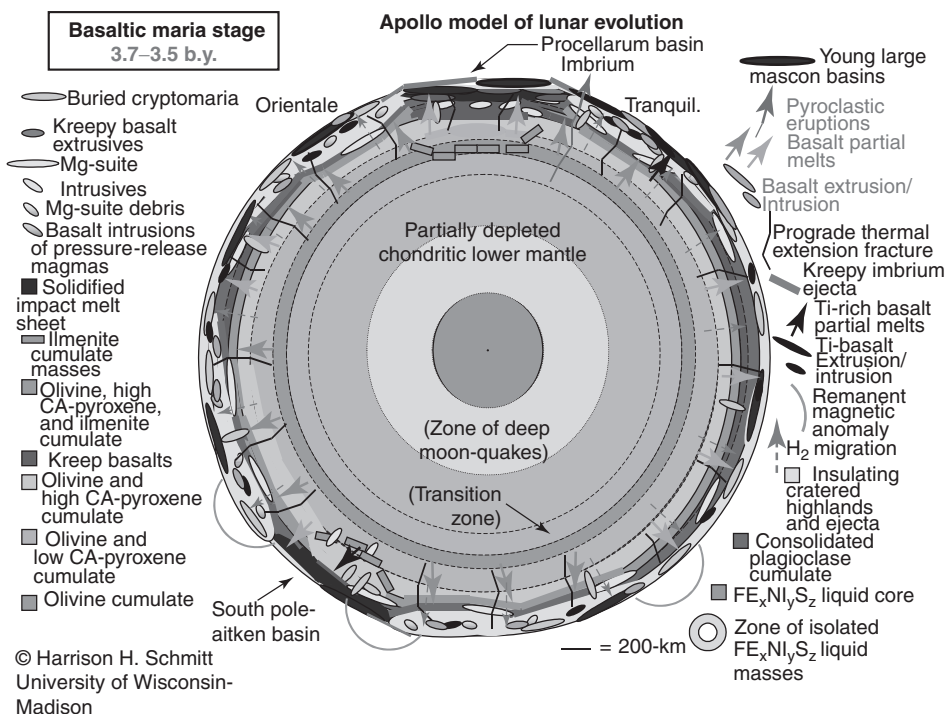


Plate 14. Apollo model of lunar evolution—Basaltic Maria Stage 3.7–3.5 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

below 400 km (476,477). These would be reflected in magmas that had progressively less titanium, less KREEP-related components, and more magnesium as deeper olivine and orthopyroxene-rich cumulates reached equilibrium with the very earliest partial melts derived from interstitial plagioclase, ilmenite, and other minor minerals. This sequence can be verified only in a general sense; the oldest samples of mapped lavas are titanium-rich (Apollos 11 and 17), and the youngest are magnesium-rich (Apollos 12 and 15 and Lunas 16 and 24). Analysis indicating that platinum group elements have higher concentrations in the older, titanium-rich basalts than in the younger, magnesium-rich basalts, is consistent with this sequence as well. Due to the early separation of iron-rich liquid from the Magma Ocean (see the previous discussion of the Magma Ocean Stage), the remaining siderophilic platinum group elements, most of which would have gone into the iron-rich liquid, would have been reconcentrated in the residual liquid from which ilmenite ultimately crystallized (478). Clearly, there are great uncertainties in attempts to correlate source region compositions and ages among various maria on the Moon. For example, in the Oceanus Procellarum maria west of Imbrium there appears to be a reversal of the sequence of old and titanium-rich to young and magnesium-rich. Many of the youngest lavas in Oceanus Procellarum that have been identified by remote sensing and dated by crater statistics (479) are also titanium-rich. This reversal may be explained by an earlier migration of dense ilmenite cumulates to deeper portions of the mantle in response to the disruption of the Procellarum basin-forming event. Such a

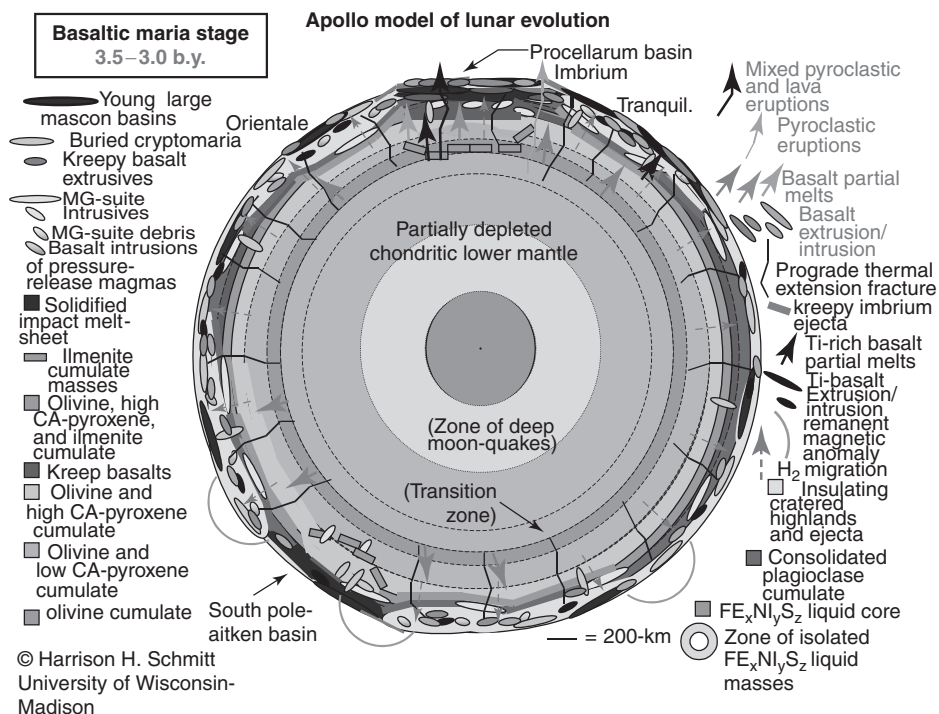


Plate 15. Apollo model of lunar evolution—Basaltic Maria Stage 3.5–3.0 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

gravitational destabilization of ilmenite cumulates has been suggested for the entire Moon (480); however, the distribution and ages of Ti-rich mare basalts are not consistent with this more global hypothesis.

Samples of basalts in the older to younger age sequence, Apollo 11 > Apollo 17 > Luna 16 > Luna 24/Apollo 15 > Apollo 12 (481–483), appear to follow this melting sequence scenario reasonably well. Greater amounts of incompatible, KREEP-related elements are also noted in older basalts (484) due to the close association of late ilmenite cumulates with the residual melt of the Magma Ocean. Although Apollo sample analyses suggest a bimodal distribution of titanium in mare basalts, remote sensing data indicate that the majority of mare surfaces show a few percent TiO₂ and smooth decreases in abundance on either side of this peak (485). Except for the old titanium-rich basalts noted, variations in major element compositions, particularly in titanium, potassium, and aluminum, can be attributed to fractional crystallization of basaltic magmas and/or melting varying proportions of ilmenite-, KREEP, and plagioclase-bearing interstitial material, respectively, in the source cumulates (486–488). Significant assimilation of crustal material is ruled out by geochemical and isotopic considerations (489). Rare-earth patterns also follow a compatible sequence; the greatest europium depletion occurred in the oldest primary mare magmas (490), magmas produced from cumulates subject to the longest interval of plagioclase crystallization. A proposed alternative scenario for producing the primary magmas for the basaltic maria through partial melting of a hybridized upper mantle

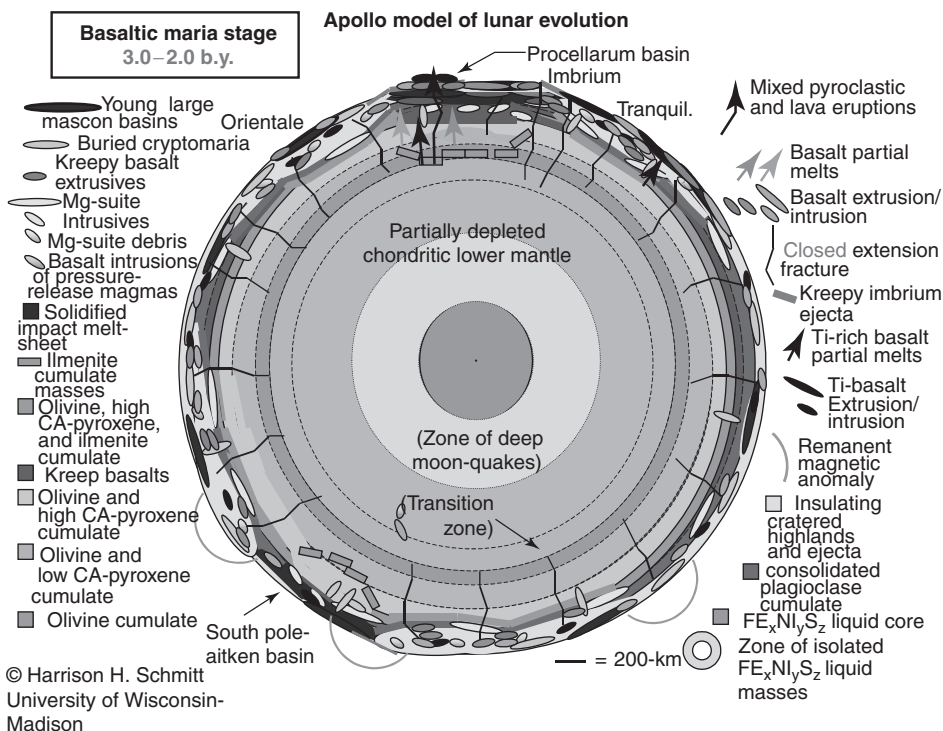


Plate 16. Apollo model of lunar evolution—Basaltic Maria Stage 3.0–2.0 b.y. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

plus assimilation of various materials (491) suffers from inconsistency with the apparent depths of partial melting as a function of time. Also, models for mantle hybridization (492) are inconsistent with the concept of the generally compositionally concentric structure of the mantle discussed before and with the improbability of overturn in the lunar mantle when fully solidified. The apparent exception to the lack of mantle hybridization, as noted before, consists of the region beneath the Procellarum Basin and possibly beneath the South Pole-Aitken Basin. Very large basin-forming events would have had the potential to destabilize the underlying mantle though pressure-release melting and may have allowed relatively dense ilmenite-rich cumulates to migrate downward.

Apollo samples and age estimates by crater statistics suggest that surface flows during the main phase of mare basalt eruption appeared first in the Serenitatis–Tranquillitatis region of the Moon. This spatial association probably exists for three reasons. First, after the formation of the Procellarum Basin and before the Serenitatis and Tranquillitatis events, as discussed previously, a thick portion of insulating megaregolith and ejecta lay over this region, accelerating reheating of the mantle. Second, due to the offset of the Moon's center of mass from its center of figure toward this region (493), a selenopotential low existed that facilitated magma access to the surface. Third, younger magmas would be forced progressively away from the area of preceding eruptions as they encountered a largely permeated and sealed crust. Although the distribution of basins

and highlands would perturb this idealized pattern, some indications of a roughly concentric color and age pattern of mare distribution around the selenopotential low have been observed (494,495). On the other hand, by 3.92 b.y. ago, the beginning of the Young Large Basin Substage, fractional remelting of the outermost cumulates of the Magma Ocean probably had progressed in many regions, if not globally, to the point of incipient eruption. The Serenitatis event (3.87 ± 0.08 b.y.) and other large events on the nearside of the Moon may have accelerated regional mantle melting by the release of lithostatic pressure resulting from the instantaneous removal of 10–20 km or more of crustal material (496). No specific evidence, however, indicates that this significantly affected the eruptive history of the maria (497). Elimination of most of the initial pressure change might have occurred through rapid rebound of the basin floors as partially fluidized mantle material moved inward and upward to compensate (498). Additionally, nearly contemporaneous movement of large fault blocks of basin rim masses into the transient crater cavity may have taken place (499).

Nonetheless, the oldest known lavas spatially associated with the Serenitatis basin crystallized about 3.82 ± 0.25 b.y. ago (500). These old, high-titanium lavas were sampled during Apollo 17 at the rim of Camelot Crater (Fig. 18). Because these mare basalts lie above Imbrium ejecta deposits and Apollo 14 and 15 samples date the Imbrium event at 3.85 ± 0.03 b.y. ago (501), the

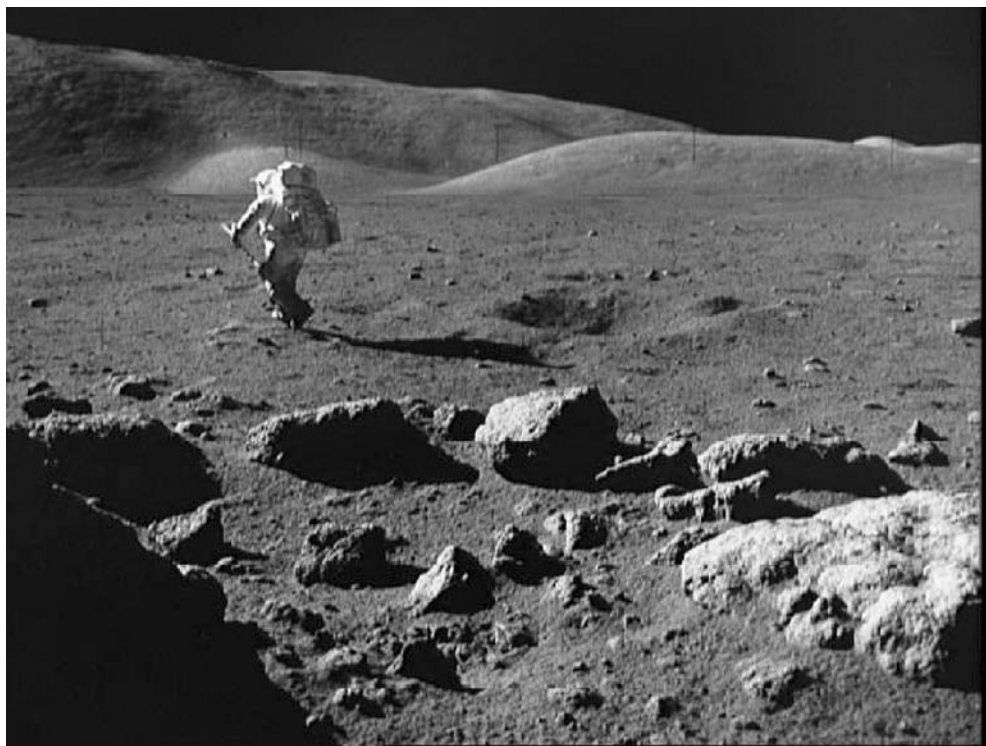


Figure 18. Mare basalt boulders at Station 5 on the rim of Camelot Crater excavated from a depth of about 150 m. The author is moving toward the Lunar Rover (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

oldest basalt ages can be constrained even more to about $3.82 \pm_{0.25}^{0.06}$ b.y. The Taurus-Littrow lava, or rather the regolith formed on them, is contiguous with a dark blue-gray annulus about 50 to 100 km wide around Serenitatis and is roughly contiguous with the titanium-rich basalts of Mare Tranquillitatus to the south. The annulus of dark blue-gray regolith underlies younger mare basalts covered with brown-gray regolith in the central portion of the southern Serenitatis Basin (502,503), regolith low in titanium, as determined by remote sensing (504). The younger, central basin lavas may be represented in the Apollo 17 sample suite by small fragments of low-titanium basalt in the regolith that have ages of 3.74 ± 0.20 b.y. (505). The most common types of titanium-rich basalts (Types A and B) sampled in the Valley of Taurus-Littrow are of a differentiated continuum between high-titanium, olivine-rich basalt and a low-potassium, olivine-poor basalt that can be produced by fractional crystallization of a single magma (506). Consideration of the depths of the craters that excavated the Type A and B samples suggests that these basalts represent most of the upper 100–150 m of the basalt section. Active seismic measurements made just after the Apollo 17 mission determined that approximately 1.175 km of basaltic rock partially filled the valley (507). *In situ* gravity measurements showing a basalt depth of ~ 1.4 km (508) roughly confirms the seismic estimate. No direct field evidence exists for the thickness of individual flows in this basalt section. The rate of extrusion may have been high enough that individual eruptions coalesced to create a few thick cooling units, possibly only one, as suggested by the chemical trends between basalt Types A and B. Field observations support this conclusion in that they disclosed only limited textural or mineralogical variability among the basalt boulders other than in vesicle concentration and grain size. The seismic profiles further suggest only two major units, one 248-m thick zone that had a seismic velocity of 250 m/s underlain by a 927-m thick zone that had a velocity of 1200 m/s, both consistent with known basalt velocities (509). The upper unit may be an intensely fractured portion of a single cooling unit because the profiles were measured largely under the highly cratered central portion of the valley. In contrast to these apparently thick Apollo 17 basalt units, samples from around craters in the basalt flows at the Apollo 12 site provided conclusive field and chemical evidence of fractional crystallization in flows only a few tens of meters thick, primarily through the settling of olivine (510,511). Olivine fractionation also seems dominant in the differentiation of the apparently much thicker Apollo 17 flows; however, ilmenite, armalcolite $[(\text{Mg,Fe})\text{Ti}_2\text{O}_5]$, and chromian spinel may be involved as well (512). On the other hand, many mare basalt flows extruded late in other eruptive areas appear to be only 10–60 m thick (513). Photographs by the Apollo 15 crew of flows exposed in Hadley Rille (514), photogeologic evidence of relatively thin flows at the surface of other basins, and the thin flow that must have protected the integrity of the orange volcanic glass deposit in Taurus-Littrow (515) indicate that this often has been the case.

Apollo 17 radar data taken from orbit (516) are consistent with a 3-km thick basalt section in the central portion of the southern Serenitatis Basin. Analysis of impact craters penetrating this mare (517) gives an estimated thickness of 4 km. A number of factors indicate that the lunar basaltic eruptions that created the maria in general occurred at high rates from many centers across wide regions. Such rapid and pervasive eruptions would result in dense plates and

possibly thick cooling units in the central portions of confining basins. For example, the innumerable deep fractures in the crust and upper mantle associated with Serenitatis and previous impacts provide a vast number of potential magma conduits. Gravity one-sixth that of Earth's and a lunar magma viscosity 10 times lower than that of the average terrestrial basaltic magma will result in a flow rate from source regions possibly 50% faster than that in Earth (518). This may lead as well to a propagation rate in the crust as much as ten times faster. The increase in volume of a partial melt compared to its crystalline equivalent would have created overpressures and/or buoyant forces (519) necessary to drive magma from the mantle and into and through the crust, accompanied by intensified microfracturing and refracturing of the lower crust. A significant likelihood also exists that the crust was in overall tension during the Basaltic Maria Stage as a result of volume increase due to thermal expansion and partial melting of the upper mantle. Such tension would have tended to keep fractures open until filled by solidified magma. All of these factors may have produced a volume of intrusive basalt dikes beneath the central portions of mare basins like Serenitatis that is great enough to contribute to the spatially associated mascon. If such a mass contribution is significant, then the degree to which modeling of mascons depends on an elevation of the crust-mantle boundary would be correspondingly reduced. The fact that the youngest flows in a given region are low volume flows can be best explained by decreasing magma production and/or the constriction or closing of most conduits by preceding eruptions. Many late eruptions, particularly of regional deposits of pyroclastic volcanic glasses, are closely associated with large, graben-forming fractures related to major basin margin structures. The very deep source regions indicated for such glasses and their associated volatiles strongly suggest that these fractures penetrate more than 500 km into the lower mantle (520), again indicating an upper mantle in structural tension at the time of eruption.

Major basalt units in the Apollo 17 suite are vesicular, as are basalts from other mare landing sites. The variability of associated mineral linings and phase equilibrium implications of these relics of an immiscible fluid phase were first studied in Apollo 11 basalts from Tranquillity Base (521). As in crystalline melt breccias, is a likely candidate for a vesicle fluid phase, although carbon monoxide is another strong possibility (522–524). Unlike the probable solar wind origin of the hydrogen in melt breccias, however, hydrogen or carbon monoxide in the mare basalt magmas would need to be derived from primordial sources (525). In the case of hydrogen, primordial water would be decomposed by downwardly migrating $\text{Fe}_x\text{Ni}_y\text{S}_z$ liquid separated from the Magma Ocean and would produce FeO and hydrogen. The total absence of any indication of water associated with the vesicle fluid phase demonstrates that all primordial water in the source materials for the Magma Ocean has been lost to space or has decomposed. Further, the lack of evidence that water was associated with the adsorbed volatiles in pyroclastic volcanic glasses indicates that no water remains in the lower, more primitive mantle and supports the hypothesis of a broadly disseminated migration of $\text{Fe}_x\text{Ni}_y\text{S}_z$ liquid through this material to form the lunar core. The presence of hydrogen as a component of lunar magmas would affect other aspects of magmatic activity, including, density, viscosity, fractional flotation of minerals adhering to vesicle walls, and the dynamics of late-stage eruptions from cooling intrusives.

Fluorine is slightly concentrated on the walls of basalt vesicles relative to the rock as a whole (526). Fluorine constitutes an important component of the volatile fraction of the orange and green volcanic glasses, so a broader attempt to characterize the vesicle fluids in lunar basalts might be of significant interest.

Regional dark mantle deposits have been mapped around the southern edge of the Serenitatis Basin (527,528) (Fig. 19) as well as in many other locations on the Moon (529). These deposits and similar pyroclastic materials sampled by Apollo 17 characterize late-stage mare volcanism at the edges of a few other large basins, including Imbrium (green volcanic glasses sampled by Apollo 15). As previously discussed, the volcanic glasses associated with these deposits and found ubiquitously as a component of mare regolith at all of the Apollo sites have



Figure 19. Regional deposit of dark mantle material near the southwestern edge of Serenitatis. Orbital observation during the Apollo 17 mission indicated that these deposits contain extensive layers of orange, red, and black glasses (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

continued to grow in significance relative to the origin, evolution, and internal structure of the Moon (530,531). The continuing interest in such glasses was catalyzed by the Apollo 17 discovery of the nearly pristine orange volcanic glass and its black, partially devitrified equivalent in the rim of Shorty Crater (532,533). Table 3 summarizes the geologic history of the area near Shorty Crater. Another type of basalt sampled in the Valley of Taurus-Littrow (Type C), a very high titanium, vesicular olivine-rich basalt, apparently formed a protective unit above the original, pre-Shorty orange and black volcanic glass deposit. Type C basalt has an unusually low Ba/Rb ratio that also distinguishes it from other basalts in the valley (534). Some such lava flow necessarily covered the deposit at Shorty soon after it was formed during a pyroclastic eruption (535) about 3.48 ± 0.03 b.y. ago (536), or about 300 m.y. after the major mare basalt eruptive activity in the valley. Orange glass argon exposure ages of 30 ± 6 m.y. (537) indicate that the original pyroclastic deposit lay unprotected on the Moon's surface for about that length of time. Further, cosmic ray track data indicate that the event that formed Shorty and exposed the orange glass occurred between about 10 ± 3 (538) and 19 m.y. ago (539). This Type C basalt made up a unit of limited lateral extent because samples of it were obtained only at Station 4 on the rim of Shorty Crater (540). No other evidence of protected orange glass was observed, even though beads of this glass are important constituents of the valley regolith (541). The Type C unit also was apparently quite thin as it, several meters of the underlying pyroclastic deposits, and a meter or so of light mantle debris (see below) must fit into a section about 14 m thick, the depth of Shorty Crater below the precrater surface. No pyroclastic material was observed on the floor of Shorty, only large blocks of fractured basalt. Most of the Type C basalt would have been pulverized and incorporated in the approximately 10 m of regolith developed in the valley during the 3.5 billion years since its eruption.

These orange pyroclastic glasses (see Figs. 6 and 7), independent of their adsorbed volatiles, appear to have originated from melting of the deep portions of the upper mantle, possibly as deep as 400 km or more below the surface (542,543). Even deeper origins are indicated for the low-titanium green glasses sampled by Apollo 15 (544). The small europium depletion in the Apollo 15 green volcanic glasses (545) suggests a deep (early) cumulate source region less affected than mare basalt source regions by the loss of europium during plagioclase separation from the Magma Ocean. On the other hand, the larger europium depletion and generally higher rare-earth concentrations in the Apollo 17 orange glasses suggest a relatively shallow (late) cumulate source region from which plagioclase has crystallized. Clearly, there is much more to be learned from the pyroclastic glasses. For example, the high-titanium orange glass turned out to be high in iron and magnesium relative to high-titanium mare basalts (546), also consistent with a source distinct from those basalts. Interpretations of analyses of trace elements (547) indicate that the source regions for some pyroclastic glasses have some chondritic characteristics with regard to these elements and include garnet as a mineral phase. These data suggest that some of the source regions of the pyroclastic glasses lie below 550 km, the apparent base, on seismic grounds, of the largely or entirely molten Magma Ocean (548,549). Further, the light gray regolith on either side of the orange glass deposit, apparently altered from the darker gray of the nearby regolith due to volatile migration, contains

Table 3. Sequence of Geologic Events in the Vicinity of Shorty Crater

Event	Thickness or dimension	Age or Δ time
Deposition of Serenitatis ejecta and faulting to form the proto-Valley of Taurus Littrow	~ 700 m (635)	3.87 ± 0.08 b.y. (636)
Development of regolith	(?)	$< \Delta 0.1$ b.y.
Deposition of Imbrium ejecta	~ 100 m (637)	3.85 ± 0.03 b.y. (638)
Development of regolith	(?)	$< \Delta 0.31$ b.y.
Eruption of basal mare basalt unit (?)	$927 \text{ m} \pm 10\%$ (639)	Between 3.88 and 3.57 b.y. (between the oldest limit on age for Imbrium and the youngest limit on oldest sampled Taurus-Littrow basalt)
Development of regolith (?)	(?)	
Eruption and <i>in situ</i> differentiation of Types A and B, high-titanium olivine basalt flows (Stations 1 and 5)	$248 \text{ m} \pm 10\%$ (640)	3.82 ± 0.05 0.25 b.y (641)
Development of regolith (74240–49? and 74260?)	A few cm	$\Delta \sim 6\text{--}41$ m.y.
Pyroclastic eruption of orange glass and black, partially devitrified glass beads (double drive tube 74001/2 and 74220)	2–4 m (Visual estimate from photographs showing a dark layer in west rim of Shorty Crater; (642)	3.48 ± 0.03 (643)
Development of regolith (no known sample)	$\sim 2\text{--}5$ cm	$\Delta \sim 20 \pm 6$ m.y. (644)
Eruption of basalt lava deposit (74245, 74255 and 74275) (Type C?) that covered the pyroclastic deposit and protected much of it from regolith formation (645) and cosmic ray exposure (646)	~ 10 m	3.45 b.y.?
Development of regolith (no known sample)	$\sim 5\text{--}10$ m (647–649)	$\Delta \sim 3.34$ b.y.
Avalanche of South Massif regolith Stations 2, 2a, and 3) to create the light mantle unit (650). Probably related to the impact of ejecta from the crater Tycho (651)	~ 1 m (at Station 4)	95 ± 6 m.y. (652)
Development of regolith (LRV-5 and 6)	$\sim 5\text{--}10$ cm (653)	$\Delta \sim 75$ m.y.
Impact creating 110-m diameter Shorty Crater and excavating to a depth of 20–25 m	~ 14 m below the Preexisting surface (654), including a few meters into the Type A/B basalt unit	10 ± 3 (655) to 19 m.y. (656)

Table 3. (Continued)

Event	Thickness or dimension	Age or Δ time
Pressure-driven eruption of gas/ bead mixtures along conduits in the radial and circumferential fractures around Shorty Crater (657) and volatile migration, particularly Na (658), into surrounding regolith or deposition of portions of orange and black pyroclastic ejecta in the rim and wall of Shorty (659)	> 70 cm	Δ seconds to minutes
Development of regolith (top of drive tube 74001)	~ 0.5 cm (660)	$\Delta 7$ –19 m.y.

unusually high Na (550), which is otherwise depleted in crustal mare basalts relative to carbonaceous chondrites (551).

After mare basalt eruptions, and associated with many if not most eruptions of regional pyroclastics (552–554), normal faulting (555) occurred in nearby mare-filled basins. These faults produced many long, often arcuate graben valleys (556) (see Fig. 19). As discussed before, this phenomenon suggests that the stress regime of the lunar crust was tensional during the Basaltic Maria Stage and for some time before. Such a tensional stress field would be expected because thermal expansion and partial melting of the upper mantle began soon after the formation of the Cratered Highland Stage's insulating megaregolith. The more slowly expanding lower mantle may have maintained this global tensional field, long after heating and partial melting of the upper mantle ceased due to mare basalt eruption and consequent heat transfer to the surface. Long duration crustal tension may explain the existence of a structural moat or trough rather than the normal talus apron between the base of the Apollo 17's South Massif and the mare basalt fill in the Valley of Taurus-Littrow (557). A sporadic but continuous widening of the fault boundary that defines the south side of the valley (558) apparently has provided space for talus accumulation. Later regional relaxation of crustal tension would have contributed to the formation of the abundant mare ridges (559). These ridges, also called wrinkle ridges, constitute common features on the exposed surfaces of the lunar maria (560). They are apparently produced by regional compressive stresses associated with eruptive episodes of basalt of an age younger than the lavas affected or with the cooling of the upper mantle after eruptive activity ceased in a given region. The east–west to north–south Jefferson-Lincoln-Lee scarp that crosses the lower slope North Massif and the Valley of Taurus-Littrow appears to be such a ridge. It probably expresses the trace of a thrust fault (561). The absence of strike-slip faults (562) in this region as well as elsewhere on the Moon probably reflects the very low shear strength of the intensely fractured and brecciated outer lunar crust. That

some shear stress can be present in the Moon is indicated by local en echelon displacement of a few graben valleys (563). The greater curvature of the Moon relative to other planets may also contribute to the lack of significant regional shear stress (564).

Mature Surface (Stage 6—Pre-Nectarian—Present). Maturation of the surface of a lunar geologic unit begins with stabilization of the unit's upper surface. Primary and secondary impacts and space radiation cause some degree of modification on every surface formed on the Moon, however transitory its exposure may be to the space environment. A quantitative maturity index has been developed (565); however, it has become increasingly clear that factors other than age can effect the measured "maturity." Although the index is useful in comparing similar units, relative iron, and titanium contents of the original materials (566), and initial particle size and volcanic glass content (567) will affect its values significantly. For most of the lunar highlands, there was a continuous process of megaregolith development followed by normal regolith development beginning with the formation of the first coherent solid surface on the Magma Ocean. At a more detailed level, regolith development began region by region, area by area with the emplacement of the last ejecta derived from a large basin-forming event or from a crater that fully penetrated older regolith. Similarly, for the lunar maria, regolith development began basin by basin with the solidification of the last mare basalt unit at the surface. Thus, the Mature Surface Stage for the lunar maria overlaps much of the Large Basin Stage on the highlands and began in a staggered sequence between 3.8 and 3.0 b.y. ago on most of the maria, depending on the age of flow surfaces.

Regolith makes up the dominant maturation product (see Figs. 4, 8, 12 and 18). This layer of pulverized debris is developed largely through impacts, the largest producing fairly sharp, irregularly cupped contact on fractured bedrock. The regolith constitutes the upper tens of meters of surface materials in the highlands; however, over the oldest of the lunar maria, such as at Apollo 11's Tranquillity Base, the regolith reaches an average maximum depth of only about 6 m. Near contacts between the maria and steep highland slopes or where pyroclastic glasses have been added to the surface of earlier formed regolith, the overall depth of largely fine-grained debris increases. This has been studied best in the Valley of Taurus-Littrow. Here, highland debris from the North and South Massifs and glass beads from pyroclastic deposits have added several meters of thickness to the regolith (568). Cosmic rays and solar wind particles also modify the lunar regolith. Cosmic rays produce a variety of spallation isotopes useful in measuring the length of time materials have been exposed at or near the lunar surface (569). They induce the production of neutrons whose energy or temperature can be measured in lunar orbit and used to determine remotely the concentrations of some interacting elements in the regolith (570). High-energy solar wind ions, largely hydrogen and helium but containing significant carbon and nitrogen and minor noble gases as well, stream continuously from the sun, guided by solar magnetic lines of force. The flux of solar wind ions at the lunar surface varies with the quantity ejected from the Sun and because of interactions with Earth's magnetosphere. Upon impinging on the lunar surface, these ions are embedded continuously in the near surface mineral and glass constituents of the regolith, are partially released later by micrometeoroid impact and diurnal

heating, and are partially retained by burial under ejecta from impacts. To some degree, released species will be entrained in the passing solar wind ("pickup ions") and are either lost entirely or reimplanted elsewhere on the Moon. For any given regolith deposit, a steady state of solar wind concentration develops that depends on the overall length of exposure, so that the measured amount of retained solar wind increases roughly with the age of the underlying unit. A definitive model of this overall process has not been published; however, the analysis of lunar soils has disclosed the approximate steady-state concentrations of solar wind volatiles in samples and cores from the various Apollo landing sites (571,572). Most sampling, bagging, transporting, splitting, and distribution of Apollo samples was not designed to prevent the loss of contained solar wind volatiles. As a consequence, the vast majority of quantitative measurements of solar wind components in the lunar regolith samples must be viewed as minimum amounts, possibly in error by as much as 50 to 100%.

Evidence is strong that solar wind hydrogen and helium are retained selectively by feldspar and ilmenite, respectively. In the earlier discussion of the origin of vesicles in crystalline melt breccias, it was noted that epithermal neutron spectra, measured by Lunar Prospector (573), indicate variations in concentrations of protons (hydrogen) in the lunar regolith. Significant concentrations by factors of 2–3 exist in the lunar maria relative to the highlands. The feldspar crystal lattice can hold hydrogen in a cation position (574); however, the specifics of retention in lunar feldspars have not yet been addressed. Hydrogen's affinity for titanium, present in the mineral ilmenite [FeTiO_3] in most maria, may contribute to the retention of solar wind hydrogen in basaltic regolith. The Prospector data also show that hydrogen, specifically protons, are concentrated by factors of 3–10 at the lunar poles, apparently related to the generally colder regolith and to approximately 20,000 km² of permanent shadow in those two areas (575,576). The temperature of surfaces in permanent shadow is a constant -230°C in contrast to the maximum daytime temperature at the equator of $+123^\circ\text{C}$ (577). The Prospector Team (578) and others using data from Clementine (579) interpreted the epithermal neutron spectra to indicate the presence of large quantities of water ice. The Clementine data have been disputed vigorously (580). Still others have suggested that most if not all the signal in the polar regions is related to solar wind hydrogen (581,582). Water ice might be deposited as a consequence of cometary impacts on the Moon, as predicted theoretically (583). The interpretation of the Lunar Prospector neutron spectrometer data indicating water appears premature, however, in the face of the proven presence of more than 100 ppm solar wind hydrogen in many soil and regolith breccia samples (584). Solar wind hydrogen would also be concentrated and preserved as a distributed regolith component in the colder polar regions as well as in permanent shadow. In permanent shadow, no significant amount of implanted solar wind hydrogen would be lost to thermal cycling and both primary and pickup ions would be continuously deposited in such areas, albeit at a slower rate due to geometric factors. Further, a continuous blanket of cometary water ice, precipitated on rare occasions in permanent shadow, would remain subject to micrometeoroid erosion comparable to that which gardens the upper few centimeters of the regolith approximately every 10 million years (see data summarized in Table 3). Unless covered by protective ejecta from impacts or deposited in one of the few extremely

deep craters near the South Pole, as now postulated by the Lunar Prospector Team (585), the water ice blanket may erode and be dispersed in a geologically short interval. Some water ice and other cometary volatiles may be preserved beneath scattered ejecta blankets or in deep craters in permanently shadowed areas, depending on the frequency of cometary impacts relative to the rate of water ice erosion. Solar wind hydrogen, however, probably accounts for most if not all of Lunar Prospector's epithermal neutron signal.

The high-titanium regolith in the Valley of Taurus-Littrow has helped to quantify the correlation of solar wind helium content with TiO_2 in Apollo samples of regolith derived primarily from mare basalt. This relationship had been well-documented in samples from other Apollo sites (586). Because ilmenite is the only significant titanium mineral in the regolith, it appears to accommodate helium in its lattice better than other minerals and glasses. Helium release signatures from lunar samples containing ilmenite have been duplicated in terrestrial ilmenite irradiated with helium ions at solar wind energies (587). Significantly, the light isotope of helium, ^3He , holds great promise as a lunar resource export to supply future fusion power plants on Earth (588,589). Although the regolith in the valley of Taurus-Littrow has concentrations of ^3He , early resource production probably would concentrate on much larger areas in the Mare Tranquillitatis (590,591). The 2000-km long volcanic province of central Oceanus Procellarum constitutes another high-titanium region of the Moon; however, until direct evaluation of the strength of the titanium-helium correlation can be made specifically for this province, Tranquillitatis probably will remain the preferred target area for early production. If it is shown that the proton signal at the lunar poles is largely the result of a 3- to 10-fold concentration of solar wind hydrogen, then helium may be concentrated there as well. Whether such a concentration advantage can offset the higher cost of mining, living, and transportation inherent in any resource recovery at the lunar poles remains to future study.

During the Mature Surface Stage, impacts continued and declined to approximately present frequency levels (592) with some recently detected variability (593). They did not, however, change the face of the Moon significantly. Three billion years ago, the full Moon would have looked very familiar to a time traveler from our day. The type example of one of the few major impacts of this stage has been that which formed the 95-km diameter crater Copernicus (594,595) about 0.85 b.y. ago (596). Secondary projectiles from another such crater, Tycho, appear to have formed the cluster of craters in the central portion of the Valley of Taurus-Littrow and triggered the avalanche or landslide that produced the light mantle deposit (597) (see Fig. 3). This association, if valid, places an age of ~ 100 m.y. on the Tycho event (598). The plume-like avalanche deposit, essentially identical mineralogically and chemically to the regolith on the slopes of the South Massif, was investigated and sampled at Stations 2, 2A, and 3. An avalanche-related origin of the light mantle deposit is supported by an apparent vertical separation of rock fragments by size within the deposit (599). *In situ* screening of near-surface debris from both South Massif talus and light-mantle, produced a significantly lower frequency of rock fragments larger than 2 cm in samples from the light mantle. Additionally, visual comparisons of the size of boulders excavated by impacts into the deposit correlated roughly with the

diameter of craters, and thus with the depth of excavation. Sorting by size and the potential availability of hydrogen released by particle interaction in the flowing feldspathic regolith further suggests that the avalanche was gas lubricated. The liberation of solar wind gases by particle abrasion in regolith samples was demonstrated during soil mechanics experiments on returned samples (600).

Possible Implications

Apollo 17 had both the privilege and the sadness to be the capstone on humankind's first venture into deep space. The privilege came in contributing to the reemphasis of the potential of free men and women when given a challenge they believe must be met. The sadness lay in not continuing to amortize the investment of human lives, families, energy, and resources that made Apollo possible. Science, however, owes much to the explorations of the Moon and became the second major objective of this technological race into the future. As a consequence of obtaining an understanding of the evolution of a second planet, we now can look at other terrestrial planets with far greater insight than ever would have been possible otherwise (Fig. 20). The record of impact activity on the Moon, particularly that between about 4.4 and 3.8 b.y. ago, represents a period of Earth history that spans the time when complex organic molecules became replicating life forms (601–603) and continents began to form. During the first 200 m.y. of this period, the Moon's dry anorthositic crust was saturated with impacts capable of forming craters 60–70 km in diameter. On Earth, the pulverized and partially vitrified crust created by a similar saturation would have continuously reacted with water to create a wide spectrum of clay species (604), whose crystal structure may have been important as templates for organic synthesis. During this same interval, two and possibly more very large impact basins formed on the Moon; their terrestrial equivalent would have been larger and had thick interior melt-sheets. The differentiation of these melt-sheets would have yielded silica-rich disks thousands of kilometers in diameter and possibly tens of kilometers thick that became potential seeds for the aggregation of early continents. The existence of a continental crust on Earth has now been placed at least at 4.3 b.y. ago (605,606) and consistent in time with the formation of very large lunar basins. The next 400 m.y. on the Moon saw about 50 large impacts that created basins at least 300 km in diameter plus many more smaller. On Earth, a much larger number of similar events may have delivered additional organic components as well as augmented the initial continental material. The global effect, however, of such highly energetic and repeated impacts potentially would have assisted and disturbed the final development of replicating life forms. In addition to formation of complex molecules in terrestrial environments (607), a continuing influx of organic chemicals may also have arrived as constituents of comets (608). Recent isotopic evidence of terrestrial biological processes about 3.8 b.y. years ago (609) is consistent with the end of large basin formation in the inner solar system. Thus, the Moon gave us a window into the first 1.5 b.y. of Earth history, a period that culminated in the first isotopic indications of biological processes by life forms on our home planet.



Figure 20. Apollo 17's view of a nearly full Earth, photographed by the author from a distance of about 34,000 miles (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Other scientific implications of lunar exploration and research relative to understanding the other terrestrial planets and the asteroids are also profound. The Moon forms one end member in the planetary mass series Earth–Venus–Mars–Mercury–Asteroids–Moon (610). Having a detailed understanding of the nature and evolution of the two end members of this series, rather than of just Earth, has increased the value of remotely sensed data about the others by orders of magnitude. For example, new questions can be asked about the sequence of events during the formation of Earth's core (611). Discrepancies between the apparent ages of Earth and the Moon (612) can be approached from new directions (613). Evidence indicating that primitive lead isotopic ratios, chondritic tungsten isotopic ratios, and increased aluminum exist in the lunar mantle below about 550 km strongly suggests that the Moon had an original chondritic core of about 1200 km in radius. Further, $^{182}\text{HF}/^{182}\text{W}$ systematics constrain the crystallization of most of the lunar Magma Ocean to within the first ~40 million years of solar system history. These data may indicate that the

Moon evolved separately from Earth and was captured rather than originating as a consequence of a giant terrestrial impact. They also suggest that the original core of Earth, as well as those of Mars, Venus, and Mercury, would have been chondritic and comparable in size to the 1200-km protocore of the Moon. Initially, cool cores would have delayed metallic core formation resulting from the downward migration of iron-nickel-sulfur liquid separates from Magma Oceans. Age correlation of the start of global magnetism on the Moon with young, large impact basins indicates a delay of about 700 million years in lunar core formation. Prelarge-basin magnetic striping detected in highland units on Mars (614) suggests a shorter, possibly ~ 300 to 400 million year delay on that planet. On Earth, the delay may have been long enough to isolate radiogenic isotopic systems temporarily from magma ocean isotopic systems in an initially chondritic core. A delay in the mixing of these two systems after core formation may explain the paradox of model radioisotopic systems giving ages for Earth that are ~ 100 million years younger than the age of the Moon and meteorites.



Figure 21. Apollo 17's view of the rising crescent Earth from behind the Moon (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The nature of Earth's crust when plate tectonic processes first began to assemble continents can be illuminated by joint consideration of the evolution of the Moon and Mars (615). Further, interpretations of spacecraft images and data from the other terrestrial planets invariably are considered first in the context of what we know about Earth and the Moon. The Asteroids, sampled by many meteorites and imaged by several spacecraft, can now be viewed confidently as largely the remains of a Moon-like planet, broken apart by interactions with Jupiter (616). Although our remotely sensed information about Mercury is limited compared to the other planets, Mariner 10 images and other data can be interpreted as disclosing a somewhat larger version of the Moon (617), but one that evolved much nearer the Sun and had a much larger metallic core. Global photography, orbiting sensors, and telerobotic landers have begun to illuminate the geologic history of Mars. Martian history now can be organized with reference to the sequence of major stages of lunar evolution, modified by our terrestrial experience with the effect of water and an atmosphere (618). Venus still holds many mysteries about its resurfaced crust in spite of the remarkable global coverage by the Magellan spacecraft's radar system (619). Using that imagery, however, and our knowledge of Earth and the Moon, we know the questions to ask about underlying materials and structures and about samples when they can be obtained from the surface of Earth's sister planet. Finally, the discovery of solar wind resources in the lunar regolith, particularly ^3He as a potential terrestrial fusion energy fuel, has joined



Figure 22. The Lunar Module Challenger at rest in the Valley of Taurus-Littrow on the Moon (courtesy of NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Earth and the Moon together as one environmental system for the future (Fig. 21). Our ability to be a truly spacefaring species in the foreseeable future may rest on the availability in space of lunar consumables, hydrogen, oxygen, water, and food. Ultimately, the returns on America's investment in Apollo (Fig. 22) will turn out to be as large and immeasurable as they were for those stimulated by Thomas Jefferson in the Louisiana Territory, Abraham Lincoln in Alaska, and many other public and private commitments to exploration and science.

ACKNOWLEDGMENTS

The author is profoundly indebted to the thousands of researchers who studied the Moon and planets through the ages and particularly since President John F. Kennedy issued his challenge for Americans to go to the Moon. The frequent citing of general review references in part acknowledges the vast numbers of individuals who contributed to any given area of investigation, and the reader is referred to these works to pursue the history of the development of particular data sets or interpretations. The author is also indebted beyond expression to the late Graham Ryder and to R. L. Korotev who provided extraordinarily valuable reviews of the technical portions of the manuscript that pertain to the origin and evolution of the Moon and were submitted for publication elsewhere. Their reviews improved the quality of those discussions immeasurably; however, the author takes full responsibility for any errors of omission or commission that remain in those or other sections.

BIBLIOGRAPHY

1. Low, G.M. Personal Notes #30. *RPI Archives and Special Collections*. Troy, NY, 1970.
2. Kennedy, J.F. *Congressional Record* - May 25, 1961, US Congress, 1961; Logsdon, J.M., *Decision to go to the Moon*. MIT Press, Cambridge, 1970.
3. Brooks, C.G. *Chariots for Apollo*. NASA, Washington, DC, 1979, p. 364.
4. Chaikin, A. *Man on the Moon*. Viking Press, 1994, pp. 389–394; Compton, W. D., *Where No Man Has Gone Before: A History of Apollo Lunar Exploration Missions*, NASA Special Publication 4214, US Government Printing Office, Washington, 1989, pp. 143–166.
5. Brooks, C.G. *Chariots for Apollo*. NASA, Washington, 1979, pp. 143–166.
6. Schmitt, H.H. In E.M. Cortright (ed.), *Apollo Expeditions to the Moon*. SP-350, NASA, Washington, 1975, pp. 265–288.
7. Hinners, N.W. *Apollo 17 Preliminary Science Report*, NASA SP-330, 1973, pp. 1-1–1-5.
8. Chaikin, A. *Man on the Moon*. Viking Press, 1994, p. 505.
9. Low, G.M. Personal Notes #30. *RPI Archives and Special Collections*, Troy, NY, 1972.
10. Johnson Space Center. *Apollo 17 Preliminary Science Report*. NASA SP-330, 1973, pp. 2-1–19-19.
11. Scott, D.H., and co-workers. *Taurus-Littrow Region—Apollo 17*. U.S. Geological Survey Misc. Geol. Invest. Map I-800, 1972, Sheet 1 of 2.
12. Wolfe, E.W., and co-workers. *Geotimes* 17 (11): 14–18 (1972).
13. Lucchitta, B.K. *Taurus-Littrow Region—Apollo 17*. U.S. Geological Survey Misc. Geol. Invest. Map I-800, 1972, Sheet 2 of 2.
14. Apollo 17 Mission Team. *Apollo 17 Mission Report*, JSC-07904. NASA, Houston, 1973, pp. 5-1 and 6-91.

15. Jones, E.M. Apollo Lunar Surf. J. <http://www.hq.nasa.gov/office/pao/History/alsj/a17/a17.html>.
16. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus–Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 1.
17. Shoemaker, E.M., and co-workers. In *Surveyor III: A Preliminary Report*. NASA Report Sp-146, 1967, pp. 9–59.
18. The Apollo 17 ALSEP actually was still providing information in 1977 when NASA deactivated it and its sister stations at four other sites.
19. Schmitt, H.H. *Science* 182: 681–690 (1973).
20. Schmitt, H.H., and E.A. Cernan. *Apollo 17 Mission Report*, JSC-07904. NASA, Houston, 1973, pp. 5-1–5-21.
21. Muehlberger, W.R., and co-workers. *Apollo 17 Mission Report*, JSC-07904. NASA, Houston, 1973, pp. 6-1 and 6-91.
22. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus–Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 280.
23. Descriptions and preliminary results of the Apollo 17 ALSEP and its experiments can be found in the *Apollo 17 Preliminary Science Report*, NASA SP-330. 1973, pp. 7-1–19-19. The detailed scientific results of experiments can be found in the *Proc. Lunar Planetary Sci. Confs.* beginning with Conference 4.
24. Pike, R.J. *Earth Planetary Sci. Lett.* 23: 265–274 (1974).
25. See summary in Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 27–53.
26. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 47–56; 62–84.
27. Croft, S.K. In P.H. Schultz and R.B. Merrill (eds), *Multi-Ring Basins, Lunar Planetary Sci. Conf. 12*, Pergamon, New York, 1981, pp. 133–148.
28. Melosh, H.J. *J. of Geophys. Res.* 87: 371–380 (1982).
29. Cintala, M.J., and R.A.F. Grieve, *Meteoritics Planetary Sci.* 33: 889–912 (1998).
30. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 151–152.
31. Ryder, G. *Lunar Planetary Sci. Conf. 30*, Abstract #1362, 1999.
32. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 285.
33. Shoemaker, E.M., and co-workers. In *Surveyor Project Final Report*, Part II, JPL Technical Report 32-1265, NASA SP-146, 1968, pp. 21–136.
34. Taylor, L.A., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1885, 1999.
35. Noble, S.K., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1810, 2000.
36. Taylor, L.A., and co-workers. *Lunar Planetary Sci. Conf. 32*, Abstract #2196, 2001.
37. Cameron, A.G.W., and W.R. Ward. *Lunar Sci. Conf. 7*, 1976, pp. 120–122.
38. Hartmann, W.H. *Origin of the Moon*. Lunar and Planetary Institute, Houston, 1986.
39. Warren, P.H. *Annu. Rev. Earth Planetary Sci.* 13: 201–240 (1991).
40. See reviews in Canup, R.M., and K. Righter. *Origin of the Earth and Moon. Part III*. University of Arizona Press, 2000, pp. 133–226.
41. Jolliff, B. L., and co-workers. *EOS* 81 (31): 349/354–355 (2000).
42. Canup, R.M., and E. Asphaug. *Nature* 412: 708–712 (2001).
43. Melosh, J. *Nature* 412: 694–695 (2001).
44. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, AGU Geophysical Monograph 95. 1996, pp. 33–46.
45. Agee, C.B. In C. B. Agee and J. Longhi (eds). *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 11–12.
46. Neal, C.R., and co-workers. *Lunar and Planetary Sci. Conf. 31*, Abstract #1944a, 2000.

47. Wood, J.A., and co-workers. *Proc. Apollo 11 Lunar Sci. Conf.* 1970, pp. 965–988.
48. Warren, P.H. *Annu. Rev. Earth Planetary Sci.* 13: 201–240 (1985).
49. Agee, C.B. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 11–12.
50. Wieczorek, M.A., and R.J. Phillips. *J. Geophys. Res.* 103: 1715–1724 (1998).
51. Warren, P.H., and J.T. Watson. *Rev. Geophys. Space Phys.* 17: 73–88 (1979).
52. Haskin, L.A., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1858, 1999.
53. Feldman, W.C., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #2056, 1999.
54. Wieczorek, M.A., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1548, 1999.
55. Korotev, R.L. *Lunar Planetary Sci. Conf. 30*, Abstract #1305, 1999a.
56. Jolliff, B.L., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1670, 1999.
57. Cumulate: An igneous rock formed by the accumulation of crystals that settle out from a magma by the action of gravity. After Glossary of Geology, American Geological Institute.
58. Parmentier, E.M., and P.C. Hess. *Lunar Planetary Sci. Conf. 30*, Abstract #1289, 1999.
59. Neal, C.R., and L.A. Taylor. Petrogenesis of mare basalts: A record of lunar volcanism. *Geochimica et Cosmochimica Acta* 56: 2177–2211 (1992).
60. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 225–228.
61. Shearer, C.K., and C. Floss. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon. Part III*, University of Arizona Press, 2000, pp. 343–344.
62. Taylor, G.J., and P.H. Warren (eds), *Workshop on Moon in Transition: Apollo 14, KREEP, and Evolved Lunar Rocks*, Technical Report Number 89-03, 1989.
63. Bell, J., and B. Hawke. *J. Geophys. Res.* 89: 6899–6910 (1984).
64. Clark, P.E., and B.R. Hawke. *Earth Moon Planets* 53: 93–107 (1991).
65. Head, J.W., and co-workers. *J. Geophys. Res.* 98: 17149–17181 (1993).
66. Williams, D.A., and co-workers. *J. Geophys. Res.* 100: 23291–23299 (1995).
67. Tera, F., and co-workers. *Earth Planetary Sci. Lett.* 22: 1–21 (1974).
68. Ryder, G. *EOS* 71: 313; 322–323 (1990).
69. Ryder, G., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon. Part III*, University of Arizona Press, 2000, pp. 475–480.
70. Ryder, G. *Lunar Planetary Sci. Conf. 32*, Abstract #1326, 2001.
71. Hartmann, W.K., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon. Part III*, University of Arizona Press, 2000, pp. 503–508.
72. Ryder, G., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon. Part III*, University of Arizona Press, 2000, pp. 475–480.
73. Cohen, B.A., T.D. Swindle, and D.A. Kring. *Science* 290: 1754–1756 (2000).
74. Spudis, P.D., and co-workers. *Science* 1848–8151 (1994).
75. Spudis, P.D. *The Once and Future Moon*. 1996, p. 164.
76. Lin, R. P., and co-workers. *Science* 281: 1481 (1998).
77. Lin, R.P., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1930, 1999.
78. Mitchell, D.L., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #2088, 2000.
79. McGetchin, T.R., and co-workers. *Basaltic Volcanism on the Terrestrial Planets*. Pergamon, New York, 1981, pp. 236–267.
80. Wilhelms, D.E. *The geologic history of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 83–104 and 227–262.
81. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1199, 1999.
82. McGetchin, T.R., and co-workers. *Basaltic Volcanism on the Terrestrial Planets*, Pergamon, New York, 1981, pp. 236–267.

83. Wilhelms, D.E., *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 83–104 and 227–262.
84. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1199, 1999.
85. Shoemaker, E.M., and co-workers. in *Surveyor Project Final Report, Part II. Science Results*. NASA, Technical Report 32-1265, 1968, pp. 58–108.
86. Schmitt, H.H. In J.B. Thompson Volume, *Am. Mineral.* 76: 773–784 (1991).
87. Schmitt, H.H. *Workshop on New Views of the Moon II*. Lunar and Planetary Institute, Contribution No. 980, 57, 1999.
88. Jones, J.H., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon, Part III*, University of Arizona Press, 2000, pp. 197–216.
89. Stewart, G.R. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon, Part III*, University of Arizona Press, 2000, pp. 217–226.
90. Pressure-release melting: denotes a process by which rocks very near to their initial melting point melt as a consequence of the removal of the pressure or weight of overlying rocks. Sometimes referred to as “decompression melting.”
91. Wilhelms, D.E. *To a Rocky Moon*. University of Arizona Press, 1993, pp. 1–171.
92. Wilhelms, D.E. The Geologic History of the Moon. US Geological Survey Professional Paper 1348, 1987, Chap. 7.
93. “Stage:” A time-stratigraphic unit next in rank below “series” and “substage”—*Glossary of Geology*, American Geological Institute.
94. Wilhelms, D.E. The Geologic History of the Moon. U.S. Geological Survey Professional Paper 1348, 1987.
95. Schmitt, H.H. *Am. Mineral.* 76: 773–784 (1991).
96. Spudis, P.D. *The Once and Future Moon*. Smithsonian, Washington, 1996, pp. 83–169.
97. Taylor, S.R. In P.R. Weissmann, and co-workers (eds), *Encyclopedia of the Solar System*, Academic Press, San Diego, 1999, pp. 247–275.
98. Subsequent portions of this work.
99. “Pre-Nectarian” and other terms in parentheses in this summary are from the accepted time-stratigraphic nomenclature of the U.S. Geological Survey and summarized by Wilhelms, D.E. The Geologic History of the Moon. U.S. Geological Survey Professional Paper 1348, 1987, Chap. 7.
100. Taylor, S.R. *Planetary Science: A Lunar Perspective*. Lunar and Planetary Institute, Houston, 1982, pp. 409–431.
101. Taylor, S.R., and T.M. East. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Monograph 95, p. 41.
102. Patterson, C. *Geochimica et Cosmochimica Acta* 10: 230–237 (1956).
103. Carlson, R.W., and G.W. Lugmair. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 25–44.
104. Alexander, C.M.O'D., and co-workers. *Science* 293: 68 (2001).
105. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, p. 156.
106. Cameron, A.G.W., and J.W. Truran. *Icarus* 30: 447–461 (1977).
107. Vanhala, H.A.T., and A.P. Boss. *Lunar Planetary Sci. Conf. 30*, Abstract #1433, 1999.
108. Jacobsen, S.B. *Lunar Planetary Sci. Conf. 30*, Abstract #1978, 1999.
109. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, p. 41.
110. Weidenschilling, S.J., and J.N. Cuzzi. In E.H. Levy and J.I. Lunine (eds), *Protostars and Planets III*, 1993, pp. 1031–1060.
111. Wetherill, G.W. *Icarus* 119: 219–238 (1996).

112. Alexander, C.M.O'D., and co-workers, *Science* 293: 64–68 (2001).
113. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, pp. 36–38.
114. Jones, J. and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 197.
115. Taylor, S.R. *Planetary Science: A Lunar Perspective*. Lunar and Planetary Institute, Houston, 1982, p. 424.
116. Sputis, P.D. *The Once and Future Moon*. Smithsonian, Washington, 1996, pp. 161–163.
117. Wiechert, U., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1669, 2000.
118. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, p. 42.
119. Jones, J. and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 197.
120. See review by Cameron, A.G.W. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon. Part III*, University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 133–144.
121. Canup, R.M., and E. Asphaug. *Nature* 412: 708–712 (2001).
122. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, pp. 34 and 43.
123. Jones, J., and H. Palme. In R.M., Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 197–216.
124. Schmitt, H.H. *Lunar Planetary Sci. Conf. 31*, Abstract #1691, 2000.
125. Jolliff, B. L., and co-workers. *EOS* 81(31): 354 (2000).
126. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 197–216.
127. Alfven, H., and G. Arrhenius. *The Moon* 5: 210–225 (1972).
128. Schmitt, H.H. *Am. Mineral.* 76: 775–776 (1991).
129. Wetherill, G.W. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, p. 75.
130. Halliday, A.N., and M.J. Drake. *Science* 283: 1861–1863 (1999).
131. Cameron, A.G.W. *Lunar Planetary Sci. Conf. 30*, Abstract #1150, 1999.
132. Agnor, C.B. and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1878, 1999.
133. Jacobsen, S.B. *Lunar Planetary Sci. Conf. 30*, Abstract #1978, 1999.
134. Stewart, G. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 217–226.
135. O'Neill, H. *Science* 292: 2016–2017 (2001).
136. Cameron, A.G.W. In R.M. Canup, and K. Righter (eds), *Origin of the Earth and Moon. Part III*, University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 133–144.
137. Mueller, S. and co-workers. *J. Geophys. Res.* 93: 6338–6352 (1988).
138. Hood, L.L., and M.T. Zuber. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 404.

139. Weidenschilling, S.J., and co-workers. In W.K. Hartmann and co-workers (eds), *Origin of the Moon*. Lunar and Planetary Institute, Houston, 1986, pp. 731–762.
140. Jones, J. and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 213–214.
141. Alfven, H., and G. Arrhenius. *The Moon* 5: 216 (1972).
142. Wetherill, G.W. *Geochimica et Cosmochimica Acta*, 58: 4513–4520 (1994).
143. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 201–204.
144. Kreutzberger, M.E., and co-workers. *Geochimica et Cosmochimica Acta* 50: 91–98 (1986).
145. O'Keefe, J.D., and T.J. Ahrens. *Proc. Lunar Sci. Conf.* 6: 2831–2844 (1975).
146. Orphal, D.L., and co-workers. *Proc. Lunar Sci. Conf.* 11: 2309–2323 (1980).
147. Humayun, M., and R.N. Clayton. *Geochimica et Cosmochimica Acta* 59: 2131–2148 (1995).
148. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, p. 41.
149. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, p. 34.
150. Walker, D., and co-workers. *Lunar Sci. Conf.* 8: 1521–1547 (1977).
151. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 198–199.
152. Schmitt, H.H. *Science* 182: 681–690 (1973).
153. Delano, J.W., and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989.
154. See Taylor, S.R. *Planetary Science: A Lunar Perspective*. Lunar and Planetary Institute, Houston, 1982, pp. 424–429.
155. Schmitt, H.H. *Lunar Planetary Sci. Conf.* 31, Abstract #1691, 2000.
156. Schmitt, H.H. *Am. Mineral.* 76: 780 (1991).
157. Schmitt, H.H. In J.W. Delano and G. H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, pp. 56–57.
158. Wasson, J.T., and co-workers. *Proc. Lunar Sci. Conf.* 7, 1976, pp. 1583–1595.
159. Meyer, C. In J.W. Delano and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, pp. 50–51.
160. Longhi, J. *Proc. Lunar Planetary Sci. Conf.* 17. *J. Geophys. Res.* 92: E349–E360 (1987).
161. Snyder, G.A., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 381.
162. Nunes, P.D., and co-workers. *Lunar Sci. Conf.* 5, 1974, pp. 1487–1514.
163. Tera, F., and G.J. Wasserberg. *Proc. Lunar Sci. Conf.* 7, 1976, p. 858.
164. Silver, L.T. personal communication.
165. Taylor, S.R. *Planetary Science*, Lunar and Planetary Institute, Houston, 1982, p. 299.
166. Snyder, G.A., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 381.

167. Lee, D., and co-workers. *Science* 278: 1098–1103 (1997).
168. Khan, A., and K. Mosegaard. *Lunar Planetary Sci. Conf. 31*, Abstract #1341, 2000.
169. Hood, L.L., and M.T. Zuber. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 397–412.
170. See Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, pp. 33–46.
171. See R.M. Canup, and K. Righter (eds), *Origin of the Earth and Moon. Part III*, University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 133–226.
172. Alexander, C.M.O'D., and co-workers. *Science* 68 (2001).
173. Lee, D., and co-workers *Science* 278: 1098–1103 (1997).
174. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 204–209.
175. Wiechert, U., and co-workers, *Lunar Planetary Sci. Conf. 31*, Abstract #1669, 2000.
176. Radiometric age determined by analysis of the total proportion of parent and daughter isotopes in a given isotopic system. After Glossary of Geology, American Geological Institute. In the case of extinct parent radioisotopes, the model age relates to the time from creation of the extinct isotope.
177. Lee, D., and co-workers. *Science* 278: 1098–1103 (1997).
178. Radiometric age determined by extrapolation of the apparent ages of distinct phases or physical components in a rock. After Glossary of Geology, American Geological Institute. In the case of extinct parent radioisotopes, the isochron age relates to the time from creation of the extinct isotope.
179. Lee, D., and co-workers. *Science* 278: 1098–1103 (1997).
180. Shearer, C.K., and H.E. Newsom. *Lunar Planetary Sci. Conf. 30*, Abstract #1362, 1999.
181. Warren, P.H. in C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 68–69.
182. Drake, M., and A. Halliday. *Origin of the Earth Conf.* Monterey, Lunar and Planetary Institute, Houston, 1998.
183. Jones, J., and H. Palme. In R.M. Canup, and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 204–209.
184. Wetherill, G.W. *Science* 228: 877–879 (1985).
185. Wetherill, G.W. *Icarus* 100: 307–325 (1992).
186. See R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon. Part II*, University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 75–132.
187. Canup, R.M., and E. Asphaug. *Nature* 412: 708–712 (2001).
188. Taylor, G.J., and M.D. Norman. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 58–65.
189. Jones, J., and H. Palme. In R.M., Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 208–209.
190. See also Palme, H. *Lunar Planetary Sci. Conf. 30*, Abstract #1763, 1999.
191. Ojakangas, G.W. *Lunar Planetary Sci. Conf. 30*, Abstract #1978, 1999.
192. Eriksson, K.A., and E.L. Simpson. *Geology* 28: 831–834 (2000).

193. Williams, J.G., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #2018, 2000.
194. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, p. 224.
195. Halekas, J.S., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1436, 2000.
196. Wood, J.A., and co-workers. *Proc. Lunar Sci. Conf. 1 1*: 965–988 (1970).
197. Smith, J.V., and co-workers. *Proc. Lunar Sci. Conf. 1 1*, 897–925 (1970).
198. Taylor, S.R., and P. Jakes. *Proc. Lunar Sci. Conf. 8*: 433–446 (1977).
199. Warren, P.H. *Annu. Rev. Earth Planetary Sci.* 13: 201–240 (1985).
200. Agee, C.B., and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991.
201. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 205–209.
202. Goins, N.R., and co-workers. *Proc. 10th Lunar Planetary Sci. Conf. 3*, 1979, pp 2421–2439.
203. Khan, A., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1341, 2000.
204. Hood, L.L., and M.T. Zuber. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 397–409.
205. Ahrens, T.J. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 13–14.
206. Srinivasan, G., and co-workers. *Science* 184: 1348–1350 (1999).
207. Wolf, R., and E. Anders. *Geochimica et Cosmochimica Acta* 44: 2111–2124 (1980).
208. Elemental and isotopic abundances in the Sun and in carbonaceous chondrites are the same except in some very special details. See Anders, E. and N. Grevesse. *Geochimica et Cosmochimica Acta* 53: 197–214 (1989).
209. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 202.
210. Lodders, K., and B. Fegley Jr. *The Planetary Scientist's Companion*, Oxford University Press, New York, 1998.
211. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, p. 38.
212. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 213.
213. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, p. 37.
214. Property of liquid phases that cannot dissolve completely in one another. After Glossary of Geology, American Geological Institute.
215. See Agee, C.B. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 11–12.
216. See Neal, C.R., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1961, 2000.
217. Elements that tend be enriched in iron-rich metallic phases produced in nature. After Glossary of Geology, American Geological Institute.
218. Taylor, S.R., and T.M. Esat. In A. Basu and S. Hart (eds), *Earth Processes: Reading the Isotopic Code*, American Geophysical Union Geophysical Monograph 95, 1996, pp. 35–36.

219. Jones, J., and H. Palme. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 209–210.
220. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, p. 188.
221. Delano, J.W. In J.W. Delano and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, pp. 28–29.
222. See the general formulation of this concept in Bowen, N.L. *The Evolution of Igneous Rocks*. Princeton University Press, Princeton, 1928.
223. Longhi, J. *Lunar Planetary Sci. Conf. 31*, Abstract #2097, 2000.
224. Shearer, C.K., and C. Floss. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 339–359.
225. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, pp. 205–208.
226. Hood, L.L., and M.T. Zuber. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 397–409.
227. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, pp. 205–208.
228. See Longhi, J. *Lunar Planetary Sci. Conf. 31*, Abstract #2097, 2000.
229. Warren, P.H., and G.W. Kallemeyn. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 72–73.
230. Snyder, G.A., and co-workers. *Geochimica et Cosmochimica Acta* 56: 3809–3823 (1992).
231. Gast, P.W., and co-workers. *Proc. Apollo 11 Lunar Sci. Conf. 2*, 1970, pp. 1143–1163.
232. Taylor, S.R. *Planetary Science: A Lunar Perspective*. 1982, p. 308, Fig. 6.32.
233. Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, p. 18.
234. Tayslor, G.J., and M.D. Norman. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report 92-03, Lunar and Planetary Institute, Houston, 1991, p. 60.
235. Taylor, S.R., and co-workers. *Proc. Lunar Sci. Conf. 4*, 1973, pp. 1448–1450.
236. Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, p. 261.
237. Green, D.H., and co-workers. *Proc. Lunar Planetary Sci. Conf. 6*, 1975, pp. 871–893.
238. Delano, J.W. *Proc. Lunar Planetary Sci. Conf. 11*, 1980, pp. 251–288.
239. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 129, Fig. 5.9.
240. Jolliff, B.L. In C. B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 28–29 and Fig. 3.
241. Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, p. 128, Fig. 5.8.
242. James, O.B., and co-workers. *Proc. Lunar Planetary Sci. Conf. 19*, 1989, p. 219–243.
243. James, O.B., and co-workers. *Proc. Lunar Planetary Sci. Conf. 21*, 1989, pp. 63–87.
244. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, pp. 220–225.
245. See review by Shearer, C.K., and C. Floss. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 344–345.
246. Taylor, G.J., and M.D. Norman. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 58–65.
247. Korotev, R.L. *Lunar Planetary Sci. Conf. 30*, Abstract #1302, 1999.

248. Wiezorek, M.A., and R.J. Phillips. *J. Geophys. Res.* 103: 1715–1724 (1998).
249. Hood, L.L., and M.T. Zuber. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 400.
250. Gabbroic and noritic: Contains clinopyroxene or orthopyroxene, respectively, greater than ~10% by volume. After Glossary of Geology, American Geological Institute.
251. Stöffler, and co-workers. In J.J. Papike and R.B. Merrill (eds), *Proc. Conf. Lunar Highlands*, Pergamon, 1980, pp. 51–70.
252. Granulite: Rock composed of even-sized, interlocking mineral grains. After Glossary of Geology, American Geological Institute.
253. Goodrich, C.A., and co-workers. *Proc. Lunar Planetary Sci. Conf. 16, J. Geophys. Res.* 89: C87–C94 (1984).
254. Lindstrom, M.M., and D.J. Lindstrom. *Proc. Lunar Planetary Sci. Conf. 16, J. Geophys. Res.* 91: D263–D276 (1986).
255. Cushing, and co-workers. *Meteoritic Planetary Sci.* 34: 185–195 (1999).
256. Korotev, R.L., and B. Joliff. *Lunar Planetary Sci. Conf. 32*, Abstract #1455, 2001.
257. Korotev, R.L., personal communication, 2001.
258. See synthesis by Spudis, P.D. *New Views of the Moon II*, 1999, p. 61.
259. See synthesis by Spudis, P.D., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1414, 2000.
260. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 225–228.
261. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 228–232.
262. Schearer, C.K., and J.J. Papike. *Lunar Planetary Sci. Conf. 31*, Abstract #1405, 2000.
263. Shearer, C.K., and C. Floss. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 343–344.
264. Heiken, G.H., and co-workers. *Lunar Sourcebook* 1991, p. 229, Table 6.9.
265. Shervais, J.W., and co-workers. *Proc. Lunar Planetary Conf. 15, J. Geophys. Res.* C25–C40 (1984).
266. Warren, P.H. *Annu. Rev. Earth Planetary Sci.* 13: 201–240 (1985).
267. Snyder, G.A., and co-workers. *Geochimica et Cosmochimica Acta* 59: 1185–1203 (1995).
268. Shervais, G.A., and J.J. McGee. *Geochimica et Cosmochimica Acta* 62: 3009–3023 (1998).
269. Jolliff, B.L. *Int. Geol. Rev.* 10: 916–935 (1998).
270. Korotev, R.L. *J. Geophys. Res.* 105: 4317–4345 (2000).
271. Pieters, C.M., and S. Tompkins. *Lunar Planetary Sci. Conf. 30*, Abstract #1286, 1999.
272. Schmitt, H.H. *Am. Mineral.* 76: 773–784 (1991).
273. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, pp. 19 and 225–229.
274. Shih, C.-Y., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1698, 2000.
275. Hess, P.C. *Lunar Planetary Sci. Conf. 31*, Abstract #1389, 2000.
276. Pieters, C.M., and S. Tompkins. *Lunar Planetary Sci. Conf. 30*, Abstract #1286, 1999.
277. Le.Mouelic, S., and co-workers. *J. Geophys. Res.* 104 (E2): 3833–3844 (1999).
278. Tompkins, S., and C.M. Pieters. *Meteorite Planetary Sci.* 34: 24–41 (1999).
279. Khan, A., and co-workers. *Geophys. Rev. Lett.* 2000, in L.L. Hood and M.A. Zuber, in R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 402, Fig. 4.
280. Korotev, R.L., personal communication, 2001.

281. Snyder, and co-workers. *Geochemica et Cosmochemica Acta* 56: 3809–3823 (1992).
282. Shearer, C.K., and C. Floss. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, p. 343.
283. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 208, Table 6.5.
284. Taylor, S.R. *Lunar Science: A Post-Apollo View*. Pergamon, New York, 1975.
285. Warren, P.H. *Annu. Rev. Earth Planetary Sci.* 13: 201–240 (1985).
286. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 190.
287. Collerson, K.D., and B.S. Kamber. *Science* 283: 1520 (1999).
288. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 462–461.
289. McKay, G.A., and L. Le. *Lunar Planetary Sci. Conf. 30*, Abstract #1286, 1999.
290. Lee, D., and co-workers. *Science* 278: 1098–1103 (1997).
291. Shearer, C.K., and C. Floss. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 354–355.
292. Spera, F.J. *Geochimica et Cosmochimica Acta* 56: 2253–2266 (1992).
293. Hess, P.C., and E.M. Parmentier. *Earth Planetary Sci. Lett.* 134: 501–514 (1995).
294. Parmentier, E.M., and P.C. Hess. *Lunar Planetary Sci. Conf. 30*, Abstract #1289, 1999.
295. Warren, P.H., and J.T. Watson. *Rev. Geophys. Space Phy.* 17: 73–88 (1979).
296. Incompatible element: elements whose ions cannot be incorporated into the crystal lattices of common rock-forming silicate minerals.
297. Warren, P.H., and G.W. Kallemeyn. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 72–73.
298. Stevenson, D.J. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report 92-03, Lunar and Planetary Institute, Houston, 1991, p. 55.
299. Longhi, J. *Lunar Planetary Sci. Conf. 31*, Abstract #2097, 2000.
300. Pritchard, M.E., and D.J. Stevenson, *Lunar Planetary Sci. Conf. 30*, Abstract #1981, 1999.
301. Lunar photogeologic mapping has identified three major volcanic centers along a roughly south to north line in western Oceanus Procellarum, including the Marius Hills, the Aristarcus Plateau, and the Rumker Hills. See Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 244.
302. Nyquest, L.E., and C.-Y. Shih. *Geochemica et Cosmochemica Acta* 56: 2213–2234 (1992).
303. Shearer, C.K., and H.E. Newsom. *Lunar Planetary Sci. Conf. 30*, Abstract #1362, 1999.
304. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 139–160.
305. Norman, M., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1552, 2000.
306. Taylor, S.R. *Planetary Science: A Post-Apollo View*. Pergamon, New York, 1982, pp. 238–240.
307. Wetherill, G.W. In C.B. Agee and J. Longhi (eds), *Workshop on the Physics and Chemistry of Magma Oceans from 1 Bar to 4 Mbar*, Technical Report Number 92-03, Lunar and Planetary Institute, Houston, 1991, pp. 74–75.
308. Alexander, C.M.O'D., and co-workers. *Science* 65 (2001).
309. Malhotra, R. *Astron. J.* 110 (1): 420–429.
310. Alfven, H., and G. Arrhenius. *The Moon* 5: 210–225 (1972).
311. Weidenschilling, S.J., and co-workers. In W.K. Hartmann and co-workers. (eds), *Origin of the Moon*. Lunar and Planetary Institute, Houston, 1986, pp. 731–762.

- 312. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 136.
- 313. Toksoz, M.N. *Ann. Rev. Earth Planetary Sci.* 2: 151–177 (1974).
- 314. Spudis, P.D. *New Views of the Moon II*. 1999, p. 61.
- 315. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 139–160.
- 316. See discussion in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 232–254.
- 317. Consortium Indomitable. *The Moon* 14: 1975.
- 318. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*, Geological Survey Professional Paper 1080, 1981, pp. 192–201.
- 319. Head, J.W., and co-workers. *J. Geophys. Res.* 98: 17149–17181 (1993).
- 320. Fisher, E.M., and C.M. Pieters. *J. Geophys. Res.* 100: 23279–23290 (1995).
- 321. Williams, D.A., and co-workers. *J. Geophys. Res.* 100: 23291–23299 (1995).
- 322. Nozette, S., and co-workers. *Science* 266: 1835–1862 (1994).
- 323. Binder, A.B., and co-workers. *Science* 281: 1475–1500 (1998).
- 324. Ryder, G. *Lunar Planetary Sci. Conf. 30*, Abstract #1362, 1999.
- 325. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey professional paper 1348, 1987, p. 145.
- 326. Cadogan, P.H. *Nature*, 250: 315–316 (1974).
- 327. Head, J.W., and co-workers. *J. Geophys. Res.* 98: 17149–17181 (1993).
- 328. Feldman, W.C., and co-workers. *Science* 281: 1489–1493 (1998); Feldman, W.C. and H. H. Schmitt, personal discussions, 2000.
- 329. Kaula, W.M., and co-workers. *Proc. 5th Lunar Sci. Conf.* 3, pp. 3049–3058.
- 330. Taylor, G.J., and P.H. Warren (eds), *Workshop on Moon in Transition: Apollo 14, KREEP, and Evolved Lunar Rocks*, Technical Report Number 89-03, 1989.
- 331. Hess, P.C. *Lunar Planetary Sci. Conf. 31*, Abstract #1389, 2000.
- 332. See Wieczorek, M.A., and R.J. Phillips. *J. Geophys.* 103: Plate 2 (1998).
- 333. See Konopliv, A.S., and co-workers. *Science* 281: 1477 (1998), Fig. 1.
- 334. Basaltic Volcanism Study Project. *Basaltic Volcanism on the Terrestrial Planets*, Pergamon, New York, 1981, pp. 950–952, Table 7.3.1.
- 335. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 208.
- 336. Huneke, J.C., and co-workers. *Earth and Planetary Sci. Lett.* 13: 375–383 (1972).
- 337. Lunatic Asylum. *Proc. Conf. Luna 24. Geochimica et Cosmochimica Acta*, Supplement 9, 657–678 (1978).
- 338. Stettler, A., and F. Albarede. *Earth and Planetary Sci. Lett.* 38: 401–406 (1978).
- 339. Schmitt, H.H. *Geology* 55–56 (1974).
- 340. Konopliv, A.S., and co-workers. *Science* 281: 1477 (1998), Fig. 1.
- 341. Wieczorek, M.A., and R.J. Phillips. *J. Geophys.* 103: Plate 2 (1998).
- 342. Lucey, P.G., and co-workers. *Science* 266: 1856 (1994), Fig. 1.
- 343. Head, J.W., and co-workers. *J. Geophys. Res.* 98: 17149–17181 (1993).
- 344. Nozette, S., and co-workers. *Science* 266: 1835–1862 (1994).
- 345. Feldman, W.C., and co-workers. *Science* 281: 1489–1493 (1998).
- 346. Blewett, D.T., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1438, 1999.
- 347. Peiters, C.M., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1438, 2000.
- 348. Blewett, D.T., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1501, 2000.
- 349. Wieczorek, M.A., and R.J. Phillips. *J. Geophys.* 103: Plate 2 (1998).
- 350. Haskin, L.A. *J. Geophys. Rev.* 103: 1679–1689 (1998).
- 351. Lawrence, D.J., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1856, 2000.
- 352. Feldman, W.C., and co-workers. *Science* 281: 1489–1493 (1998).
- 353. Blewett, D.T., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1438, 1999.

354. Feldman, W.C., and co-workers. *Science* 281: 1489–1493 (1998).
355. Khan, A., and K. Mosegaard. *Lunar Planetary Sci. Conf. 30*, Abstract #1259, 1999.
356. Lawrence, D.J., and co-workers. *Science* 281: 1484–1485 (1998).
357. Warren, P.H., and G.W. Kallemeyn. In B.L. Joliff and G. Ryder (eds), *Workshop on New Views of the Moon: Integrated Remotely Sensed, Geophysical, and Sample Data*, LPI Contribution No. 958, Lunar and Planetary Institute, Houston, 1998, pp. 75–76.
358. Schmitt, H.H. *New Views of the Moon II*, Abstract #1961, 1999.
359. Schmitt, H.H. *Lunar Planetary Sci. Conf. 30*, Abstract #1691, 2000.
360. Lawrence, D.J., and co-workers. *Science* 281: 1484–1485 (1998).
361. Haskin, L.A., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1858, 1999.
362. Feldman, W.C., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #2056, 1999.
363. Wieczorek, M.A., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1548, 1999.
364. Korotev, R.L. *Lunar Planetary Sci. Conf. 30*, Abstract #1305, 1999.
365. Jolliff, B.L., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1670, 1999.
366. Korotev, R.L. *Lunar Planetary Sci. Conf. 30*, Abstract #1305, 1999.
367. Jolliff, B.L., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1670, 1999.
368. Parmentier, E.M., and P.C. Hess. *Lunar Planetary Sci. Conf. 30*, Abstract #1289, 1999.
369. Parmentier, E.M., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1614, 2000.
370. Latham, G.V., and co-workers. *Proc. Apollo 11 Lunar Sci. Conf.* 1970, pp. 2309–2320.
371. Toksöz, M.N., and co-workers. *Proc. Lunar Sci. Conf. 3*, 3, 1972, p. 2542.
372. Wieczorek, M.A., and R.J. Phillips. *Lunar Planetary Sci. Conf. 30*, Abstract #1362, 1999.
373. Blewett, D.T., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1438, 1999.
374. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 245.
375. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 57–82.
376. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 65.
377. Wieczorek, M.A., and R.J. Phillips. *J. Geophys.* 103, Plate 2 (1998).
378. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 157.
379. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 136.
380. Peterson, C.A., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1624, 1999.
381. Compare Konopliv, A.S., and co-workers. *Science* 281: (1998), Fig. 1, p. 1477 with Wieczorek, M.A., and R.J. Phillips. *J. Geophys.* 103: Plate 2 (1998).
382. Schmitt, H.H., *Abstracts with Programs*, Geol. Soc. Am. Annu. Meet, 1999, p. A-44.
383. Schmitt, H.H. *Lunar Planetary Sci. Conf. 31*, Abstract #1821, 2000.
384. Ryder, G., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 475–492.
385. Hartmann, W.K., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 493–512.
386. Morbidelli, A. *Science* 280: 2071–2073 (1998); Murray, N., and M. Holman. *Science* 283: 1877–1881 (1999).
387. Fernandez, J.A. In P.R. Weissman, and co-workers (eds), *Encyclopedia of the Solar System*, Academic Press, San Diego, 1999, pp. 554–556.

388. Alfvén, H., and G. Arrhenius. *The Moon* 5: 210–225 (1972).
389. Cassen, P., and D.S. Woolum. In P.R. Weissman and co-workers. (eds), *Encyclopedia of the Solar System*, Academic Press, San Diego, 1999, p. 38.
390. Davis, D.R., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1624, 1999.
391. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, Fig. 5.22 and Plates 5 and 6.
392. Schmitt, H.H. In G.J. Taylor and P.H. Warren (eds), *Workshop on Moon in Transition: Apollo 14, KREEP, and Evolved Lunar Rocks*, Technical Report #89-03, Lunar and Planetary Institute, Houston, 1989, pp. 111–112.
393. Muller, P.M., and W.L. Sjogren. *Science* 161: 680–684 (1968).
394. Konopliv, A.S., and co-workers. *Science* 281: 1476–1480 (1998).
395. Keifer, W.S. *Lunar Planetary Sci. Conf. 30*, Abstract #1995, 1999.
396. Compare Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 190 with Konopliv, A. S., and co-workers. *Science* 281: Fig. 1, 1477 (1998).
397. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 170.
398. Norman, M., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1552, 2000.
399. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 163–170.
400. Korotev, R.L. personal communication, 2001.
401. See summary by Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, pp. 216–220.
402. Basaltic Volcanism Study Project. *Basaltic Volcanism on the Terrestrial Planets*. Pergamon, New York, 1981, p. 957.
403. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 224.
404. Hawke, B.R., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1956, 1999.
405. Bell, J., and B. Hawke. *J. Geophys. Res.* 89: 6899–6910 (1984).
406. Clark, P.E., and B.R. Hawke. *Earth Moon Planets* 53: 93–107 (1991).
407. Head, J.W., and co-workers. *J. Geophys. Res.* 98: 17165–17169 (1993).
408. Williams, D.A., and co-workers. *J. Geophys. Res.* 100: 23291–23299 (1995).
409. Antonenko, I. *Lunar Planetary Sci. Conf. 30*, Abstract #1703, 1999.
410. Head, J.W. Abstract in *Origins of Mare Basalts and Their Implications for Lunar Evolution*, Lunar Science Institute, 1975, pp. 61–65.
411. Taylor, L.A., and co-workers. *Earth Planetary Sci. Lett.* 66: 33–47 (1983).
412. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 209.
413. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 218–219.
414. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 188–190.
415. Wilhelms, D.E. *The Geologic History of the Moon*. US Geological Survey Professional Paper 1348, 1987, pp. 21; 216–224.
416. Koehler, U., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1822, 2000.
417. Taylor, G.J., and P.H. Warren. (eds), *Workshop on Moon in Transition: Apollo 14, KREEP, and Evolved Lunar Rocks*, Technical Report Number 89-03, 1989.
418. Shih, C.-Y., and L.E. Nyquist. In G.J. Taylor, and P.H. Warren (eds), *Workshop on Moon in Transition: Apollo 14, KREEP, and Evolved Lunar Rocks*, Technical Report Number 89-03, 1989, pp. 128–136.
419. Consortium Indomitable, *The Moon* 14: 1975.
420. Gray, C.M., and co-workers. *Lunar Science Institute Contribution*, 211D, 1974, pp. VII-1–VII-10.
421. Wilhelms, D.E. *The Geologic History of the Moon*. US Geological Survey Professional Paper 1348, 1987, Chaps 9 and 10.

- 422. Norman, M., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1552, 2000.
- 423. Tera, F., and co-workers. *Earth Planetary Sci. Lett.* 22: 1–21 (1974).
- 424. Ryder, G. *EOS* 71: 313 and 322–323 (1990).
- 425. Ryder, G., and co-workers. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 475–492.
- 426. McCauley, J.F. In Wilhelms, D.E., *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 66–76.
- 427. Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, p. 78.
- 428. James, O.B. *Proc. Lunar Planetary Sci. Conf.* 12, 1981, pp. 209–233.
- 429. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 64–65.
- 430. Settle, M., and co-workers. *Earth Planetary Sci. Lett.* 271–274 (1974).
- 431. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*, Geological Survey Professional Paper 1080, 1981, p. 119.
- 432. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 177, Table 9.2.
- 433. Crystals of one mineral irregularly scattered in a larger crystal of a second mineral. After Glossary of Geology, American Geological Institute.
- 434. See summary by Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 188–202.
- 435. Schmitt, H.H. *Science* 182: 692 (1973).
- 436. Consortium Indomitable. *The Moon*. 14 (1975).
- 437. Settle, M., and co-workers. *Earth Planetary Sci. Lett.* 271–274 (1974).
- 438. Consortium Indomitable. *The Moon* 14: 1975.
- 439. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987.
- 440. Schmitt, H.H. *The Moon* 14: 500 (1975).
- 441. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1349, 1987, p. 177, Table 9.2, samples 76275, 76295, and 76315.
- 442. Schmitt, H.H. *Science* 182: 161 (1973).
- 443. Snyder, C.W., and M. Neugebauer. *Space Res.* 4: 89–113 (1964).
- 444. Briceño, C. and co-workers. *Science* 291: 93 (2001).
- 445. Alexander, C.M.O'D., and co-workers. *Science* 293: 64 (2001).
- 446. Behrens, H., and G. Juller. *Mineral Mag.* 59: 15–24 (1995).
- 447. Feldman, W.C., and co-workers. *Science* 281: 1496–1500 (1998).
- 448. Schmitt, H.H., and co-workers. *SPACE 2000, Proc. Conf. Am. Soc. Civil Eng.* 2000.
- 449. Gault, D.E., and E.D. Heitowit. *Proc. Symp. Hypervelocity Impact 6*, Cleveland, 1963, 2, pp. 420–456.
- 450. Anders, E., and co-workers. *The Moon* 8: 3–24.
- 451. Ryder, G. *Lunar Planetary Sci. Conf. 30*, Abstract #1362, 1999.
- 452. See Melosh, H.J. *Lunar Planetary Sci. Conf. 31*, Abstract #1903, 2000.
- 453. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 76 and 215.
- 454. Lin, R.P., and co-workers. *Science* 281: 1481 (1998).
- 455. Lin, R.P., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1930, 1999.
- 456. Mitchell, D.L., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #2088, 2000.
- 457. Kaydash, V.G., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1044, 1999.
- 458. Hood, L.L., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1251, 2000.
- 459. Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, p. 209.
- 460. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1278, 2000.

461. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, Chap. 5.
462. McGetchin, T.R., and co-workers. *Basaltic Volcanism on the Terrestrial Planets*. Pergamon, New York, 1981, pp. 236 and 752–753.
463. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1199, 1999.
464. Head, J.W. Abstract in *Origins of Mare Basalts and their Implications for Lunar Evolution*. Lunar Science Institute, 1975, pp. 61–65.
465. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1199, 1999.
466. See summary in Basaltic Volcanism Study Project. *Basaltic Volcanism on the Terrestrial Planets*. Pergamon, New York, 1981, pp. 236–267.
467. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 186–192.
468. Longhi, J. *Geochimica et Cosmochimica Acta* 56: 2235–2251 (1992).
469. Nyquist, L.E., and C.-Y. Shih. *Geochimica et Cosmochimica Acta* 56: 2213–2234 (1992).
470. James, O.B., and T.L. Wright. *Geol. Soc. Am. Bull.* 83: 2357–2382 (1972).
471. Schmitt, H.H., and co-workers. *Proc. Apollo 11 Lunar Sci. Conf.*, 1, 1970, pp. 11–13.
472. Schmitt, H.H., and B.L. Sutton, *Abstracts 2nd Lunar Sci. Conf.*, 1971.
473. See summary in Basaltic Volcanism Study Project. *Basaltic Volcanism on the Terrestrial Planets*. Pergamon, New York, 1981, pp. 236–267.
474. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 205–208.
475. James, O.B., and T.L. Wright. *Geo. Soc. Am. Bull.* 83: 2357–2382 (1972).
476. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 208, Table 6.5.
477. Schmitt, H.H. *Space Sci. Rev.* 18: 267 (1975).
478. Neal, C.R., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1961, 2000.
479. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1278, 2000.
480. Parmentier, E.M., and P.C. Hess. *Lunar Planetary Sci. Conf. 30*, Abstract #1289, 1999.
481. Basaltic Volcanism Study Project. *Basaltic Volcanism on the Terrestrial Planets*. Pergamon, New York, 1981, pp. 950–952, Table 7.3.1.
482. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 186–192 and 208–209.
483. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1199, 1999.
484. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 261–262.
485. Giguere, T.A., and co-workers. *Lunar Planetary Science Conf. 30*, Abstract #1465, 1999.
486. Nyquist, L.E., and co-workers. *Earth Planetary Sci. Lett.* 55: 335–355 (1981).
487. Longhi, J. *Geochimica et Cosmochimica Acta* 56: 2235–2251 (1992).
488. Shearer, C.K., and C. Floss. In R.M. Canup and K. Righter (eds), *Origin of the Earth and Moon*. University of Arizona Press, Tucson, and Lunar and Planetary Institute, Houston, 2000, pp. 350–355.
489. Neal, C.R., and L.A. Taylor. *Geochimica et Cosmochimica Acta* 56: 2177–2211 (1992).
490. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 190–191.
491. Shearer, C.K., and J.J. Papike. *Lunar Planetary Sci. Conf. 30*, Abstract #1365, 1999.
492. Spohn, T., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1768, 1999.
493. Kaula, W.M., and co-workers. *Proc. 5th Lunar Sci. Conf.*, 3, pp. 3049–3058.
494. Schmitt, H.H. *Geology* 2: 55–56 (1974).
495. Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1199, 1999.
496. Head, J.W., and co-workers. *Proc. Lunar Sci. Conf. 6*, 1975, pp. 2805–2830.
497. See Hiesinger, H., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1199, 1199.

498. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 79–81.
499. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 65–70.
500. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 205.
501. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 212.
502. Schmitt, H.H. *Geology* 2: 55–56 (1973).
503. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 94.
504. Elphic, R.C., and co-workers. *Science* 281: 1493–1496.
505. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. Geological Survey Professional Paper 1080, 1981, p. 205.
506. See summary in Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. Geological Survey Professional Paper 1080, 1981, pp. 203–205.
507. Kovach, R.L., and co-workers. *Apollo 17 Preliminary Science Report*. 1973, pp. 10-1–10-12.
508. Cooper, M.R., and co-workers. *Rev. Geophys. Space Phys.* 12: 291–308 (1974).
509. Kovach, R.L., and co-workers. *Apollo 17 Preliminary Science Report*. 1973, p. 10-1.
510. James, O.B., and T.L. Wright. *Geol. Soc. Am. Bull.* 83: 2357–2382 (1972).
511. Schmitt, H.H., and B.L. Sutton. *Abstracts of the 2nd Lunar Sci. Conf.*, 1971.
512. Rhodes, J.M., and co-workers. *Proc. Lunar Sci. Conf. 5, 2, Geochimica et Cosmochimica Acta*, Supplement 5: 1097–1117 (1976).
513. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 95–99.
514. Howard, K.A., and co-workers. *Proc. Lunar Sci. Conf. 3, 1*, 1972, pp. 10–13.
515. Schmitt, H.H. In J.W. Delano and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, pp. 56–57.
516. Peeples, W.J., and co-workers. *J. Geophys. Res.* 83: 3459–3468.
517. Solomon, S.C., and J.W. Head. *Rev. Geophys. Space Phys.* 18: 107–141 (1980).
518. Taylor, G.J. In J.W. Delano and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, p. 68.
519. Wieczorek, M.A., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1520, 2000.
520. See also Arkani-Hamed, J. In J.W. Delano and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, pp. 15–19.
521. Schmitt, H.H., and co-workers. *Proc. Apollo 11 Lunar Sci. Conf. 1*, 1970, pp. 11–13.
522. Sato, M. *Proc. Lunar Sci. Conf. 7*, 1976, pp. 1323–1344.
523. Housley, R.M. *Proc. Lunar Sci. Conf. 9*, 1978, pp. 1473–1484.
524. Sato, M., *Proc. Lunar Planetary Sci. Conf. 10*, 1979, pp. 311–325.
525. Schmitt, H.H. *Workshop on New Views of the Moon II*, Lunar and Planetary Institute Contribution No. 980, 1999, p. 57.
526. Goldberg, R.H., and co-workers. *Proc. 7th Lunar Sci. Conf. 7*, 1976, pp. 1597–1613.
527. Carr, M.H. *U.S. Geological Survey Map I-489 (LAC-42)*, 1966, scale 1:1,000,000.
528. Lucchitta, B.K., and H.H. Schmitt. *Proc. Lunar Sci. Conf. 5, 3*, 1974, pp. 2427–2441.
529. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional paper 1348, 1987, p. 234, Plate 4A.

530. Delano, J.W. and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989.
531. Gaddis, L.R., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1732, 1999.
532. Schmitt, H.H. *Science* 182: 687–688 (1973).
533. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. Geological Survey Professional Paper 1080, 1981, pp. 205–208.
534. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*, Geological Survey Professional Paper 1080, 1981, p. 204.
535. Schmitt, H.H. *Science* 182: 780 (1973).
536. Tera, F., and G.J. Wasserberg. *Lunar Science VII, Abstracts of the 7th Lunar Sci. Conf.* Lunar Science Institute, Houston, 1976, pp. 858–860.
537. Summarized in Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 101.
538. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 96–98.
539. Eugster, O. *Proc. 8th Lunar Sci. Conf. 3*, 1977, pp. 3059–3082.
540. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 103 and 202–205.
541. Heiken, G.H., and co-workers. *Lunar Sourcebook*, 1991, p. 289.
542. Green, D.H., and co-workers. *Proc. Lunar Planetary Sci. Conf. 6*, 1975, pp. 871–893.
543. See summary in Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 208.
544. Delano, J.W. *Proc. Lunar Planetary Sci. Conf. 11*, 1980, pp. 251–288.
545. See summary in Basaltic Volcanism Study Project, *Basaltic Volcanism on the Terrestrial Planets*, Pergamon, New York, 1981, p. 248.
546. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 261.
547. Neal, C.R., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1944, 2000.
548. Khan, A., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1341, 2000.
549. Hood, L.L. *Lunar Planetary Sci. Conf. 31*, Abstract #1249, 2000.
550. Korotev, R.L., and D.T. Kremser. *Lunar Planetary Sci. Conf. 221*, 1992, pp. 275–301.
551. Taylor, S.R. *Planetary Science*. Lunar and Planetary Institute, Houston, 1982, p. 306.
552. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 86–93.
553. Lucchitta, B.K., and H.H. Schmitt. *Proc. Lunar Sci. Conf. 5 3*, 1974, pp. 2427–2441.
554. Petrycki, J.A., and L. Wilson. *Lunar Planetary Sci. Conf. 30*, Abstract #1335, 1999.
555. A steeply dipping fault in which the upper wall appears to have moved downward. After Glossary of Geology, American Geological Institute.
556. A valley bounded by normal faults. After Glossary of Geology, American Geological Institute.
557. Schmitt, H.H. *Science* 182: 683 (1973).
558. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 213 and Plate 1.
559. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 107–110.
560. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 107–111.

561. A shallowly dipping fault in which the upper surface appears to have moved over the lower surface.
562. A steeply dipping fault in which one surface appears to have moved horizontally against the other.
563. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 118.
564. Freed, A.M., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1691, 1999.
565. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 317–321.
566. Staid, M.I., and C.M. Pieters. *Lunar Planetary Sci. Conf. 30*, Abstract #1724, 1999.
567. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 208.
568. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 103 and 207.
569. Lipschutz, M.E., and L. Schultz. In P.R. Weissman, and co-workers (eds), *Encyclopedia of the Solar System*. Academic Press, San Diego, 1999, pp. 664–665.
570. Feldman, W.C., and co-workers. *Science* 281: 1489–1500 (1998).
571. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 354–356 and 436–448.
572. Fegley, B. Jr., and T.D. Swindle. Lunar volatiles: Implications for lunar resource utilization. In J. Lewis and co-workers (eds), *Resources of Near-Earth Space*. University of Arizona Press, 1993, pp. 393–395.
573. Feldman, W.C., and co-workers. *Science* 281: 1496–1500 (1998).
574. Behrens, H., and G. Juller. *Miner. Mag.* 59: 15–24 (1995).
575. Margo, J.L., and D.B. Campbell. *Lunar Planetary Sci. Conf. 30*, Abstract #1897, 1999.
576. Bussey, D.B.J., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1731, 1999.
577. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 36, Table 3.2.
578. Feldman, W.C., and co-workers. *Science* 281: 1496–1500 (1998).
579. Nozette, S., and co-workers. *Lunar Planetary Sci. Conf. 30*, Abstract #1665, 1999.
580. Simpson, R.A., and G.L. Tyler. *J. Geophys. Res.* 104 (E2): 3845–3862 (1999).
581. Schmitt, H.H. *Workshop on New Views of the Moon II*, Lunar and Planetary Institute Contribution No. 980, p. 57.
582. Schmitt, H.H., and co-workers. *SPACE 2000, Proc. Conf. Am. Soc. Civil Eng.*
583. Watson, K., and co-workers. *J. Geophys. Res.* 66: 3033 (1961).
584. Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, p. 444.
585. Feldman, W.C., and co-workers. *SPACE 2000, Proc. Conf. Am. Soc. Civil. Eng.*
586. Cameron, E.N. *2nd Conf. Lunar Bases Space Activities 21st Century*, Lunar and Planetary Institute, Houston, 1988, pp. 89–97.
587. Harris-Kuhlmann, K.R. *Trapping and diffusion of helium in lunar materials*, Ph.D. Dissertation, University of Wisconsin-Madison, 1998.
588. Wittenberg, L.J., and co-workers. *Fusion Technol.* 10: 167–178 (1986).
589. Schmitt, H.H. *J. Aerosp. Eng.* 60–67 (April 1997).
590. Cameron, E.N. *Evaluation of the Regolith of Mare Tranquillitatis as a Source of Volatile Elements*, Technical Report, WCSAR-TR-AR3-9301-1, 1993.
591. Johnson, J.R. *Geophys. Res. Lett.* 26 (3): 385–388 (1999).
592. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, Fig. 7.16.
593. Jolliff, B.L., and co-workers. *EOS* 81 (31): 354 (2000).
594. Shoemaker, E.M. In Z. Kopal (eds), *Physics and Astronomy of the Moon*. 1962, pp. 283–359.
595. Schmitt, H.H., and co-workers. *Geological Map of the Copernicus Quadrangle of the Moon*. U.S. Geological Survey Geological Atlas of the Moon. I-515 (LAC 58), 1967.

596. Silver, L.T. *EOS* 52: 534 (1971).
597. Lucchitta, B.K. *Icarus* 30: 80–96 (1977); Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 209–213.
598. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 213.
599. Schmitt, H.H. *Science* 184: 686 (1973).
600. Carrier, W.D., III, and co-workers. *J. Soil Mech. Found. Div. Am. Soc. Civil Eng.* 99: 979–996 (1973).
601. Schmitt, H.H. *Abstracts with Programs, Geol. Soc. Am. Annu. Meet.* 1999, p. A-44.
602. Mojzsis, S.J., and T.M. Harrison. *GSA Today* 10 (4): 1–6 (2000).
603. Kring, D.A. *GSA Today* 10 (8): 1–7 (2000).
604. Ferris, J.P., and co-workers. *Nature* 384: (1996).
605. Scherer, E., and co-workers. *Science* 293: 686 (2001).
606. Kramer, J. *Science* 293: 619–620 (2001).
607. Huber, C., and G. Wachtershauser. *Science* 281: 670–672 (1998).
608. Mumma, J.M., and co-workers. *Science* 272: 1310–1314 (1996).
609. Mojzsis, S.J., and co-workers. *Nature* 384: 55 (1996).
610. See relevant chapters in this *Encyclopedia* and in P.R. Weissman and co-workers. *Encyclopedia of the Solar System*. Academic Press, San Diego, 1999.
611. Halliday, A.N., and M.J. Drake. *Science* 283: 1861–1863 (1999).
612. Podosek, F.A. *Science* 283: 1863–1864 (1999).
613. Schmitt, H.H. *Abstracts with Programs, Geol. Soc. Am. Annu. Meet.*, in press.
614. Connerney, J.E.P., and co-workers. *Science* 284: 794–798.
615. Acuña, M.H., and co-workers. *Science* 284: 790–798 (1999).
616. This *Encyclopedia* and Britt, D.T. and L.A. Lebofsky. In P.R. Weissman, and co-workers. *Encyclopedia of the Solar System*. Academic Press, San Diego, 1999, pp. 585–606.
617. This *Encyclopedia* and Strom, R.G. Mercury. In Weissman, P.R., and co-workers *Encyclopedia of the Solar System*. Academic Press, San Diego, 1999, pp. 123–146.
618. Carr, M.H. Mars. In M.H. Carr and co-workers. *The Geology of the Terrestrial Planets*, NASA SP-469, NASA, Washington, 1984, pp. 207–263.
619. This *Encyclopedia* and Head, J.W., and A. Basilevski. Venus: Surface and interior. In P.R. Weissman and co-workers. *Encyclopedia of the Solar System*, Academic Press, San Diego, 1999, pp. 161–190.
620. See Heiken, G.H., and co-workers. *Lunar Sourcebook*. 1991, pp. 65–73.
621. Settle, M., and co-workers. *Earth Planetary Sci. Lett.* 271–274 (1974).
622. Consortium Indomitable. *The Moon* 14: (1975).
623. Kovach, R.L., and co-workers. *Apollo 17 Preliminary Science Report*, NASA SP-330, 1993, p. 10-1.
624. Wilhelms, D.E. *The Geologic History of the Moon*. US Geological Survey Professional Paper 1348, 1987, p. 224.
625. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 119–145.
626. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 177, Table 9.2.
627. Settle, M., and co-workers. *Earth Planetary Sci. Lett.* 271–274 (1974)
628. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 119–145.

629. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 177, Table 9.2, samples 76015, 76215, 77075, 77115, and 77135.
630. Settle, M., and co-workers. *Earth Planetary Sci. Lett.* 271–274 (1974).
631. Schmitt, H.H. *The Moon* 14: 500 (1975).
632. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 177, Table 9.2, samples 76275, 76295, and 76315.
633. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, pp. 103–144.
634. Schmitt, H.H. *Am. Mineral.* 76: 776 (1991).
635. Settle, M., and co-workers. *Earth Planetary Sci. Lett.* 271–274 (1974).
636. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 177, Table 9.2, samples 76015, 76215, 77075, 77115, and 77135.
637. Settle, M., and co-workers. *Earth Planetary Sci. Lett.* 271–274 (1974).
638. Wilhelms, D.E. *The Geologic History of the Moon*. U.S. Geological Survey Professional Paper 1348, 1987, p. 212.
639. Kovach, R.L., and co-workers. *Apollo 17 Preliminary Science Report*, 1973, p. 10-1.
640. Kovach, R.L., and co-workers. *Apollo 17 Preliminary Science Report*, 1973, p. 10-1.
641. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 205; this work.
642. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, pp. 205–208.
643. Tera, F., and G.J. Wasserberg. *Lunar Sci.* VII, Abstracts, Lunar Science Institute, Houston, 1976, pp. 858–860.
644. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 101.
645. Basu, A., and co-workers. In J.W. Delano and G.H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, pp. 20–21.
646. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 101.
647. Muehlberger, W.R., and co-workers. *Apollo 17 Preliminary Science Report*, 1973, pp. 6–51.
648. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 207.
649. Cooper, and co-workers. *Rev. Geophys. Space Phys.* 12: 291–308 (1974).
650. Schmitt, H.H. *Science* 182: 686 (1973).
651. Lucchitta, B.K. *Icarus* 30: 80–96 (1977).
652. Arvidson, R., and co-workers. *Lunar Sci.* VII, Abstracts, Lunar Science Institute, Houston, 1976, pp. 25–27.
653. Schmitt, H.H. *Science* 182: 690 (1973).
654. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 207.
655. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 101.
656. Eugster, O. *Proc. Lunar Sci. Conf.* 8, 3: 3059–3082 (1977).

657. Schmitt, H. H. In J.W. Delano and G. H. Heiken (eds), *Workshop on Lunar Volcanic Glasses: Scientific and Resource Potential*, Technical Report #90-02, Lunar and Planetary Institute, 1989, pp. 56–57.
658. Korotev, R.L., and co-workers. *Lunar Planetary Sci. Conf. 31*, Abstract #1363, 2000.
659. Wolfe, E.W., and co-workers. *The Geologic Investigation of the Taurus-Littrow Valley: Apollo 17 Landing Site*. U.S. Geological Survey Professional Paper 1080, 1981, p. 98.
660. Schmitt, H.H. *Science* 182: 5–16 (1973).

H.H. SCHMITT

University of Wisconsin—Madison
Wisconsin

ARIANE ROCKET PROGRAM

This article describes the circumstances that led Europe to go ahead with the Ariane Program in 1973 and subsequently to decide on a series of follow-up versions, up to and including Ariane 5. The article also describes the main management principles adopted that have varied little in 20 years and includes brief details of the successive launcher configurations from Ariane 1 to Ariane 5.

Page of History

In 1972, the European space community was in a state of crisis. It is interesting to look back at the situation which preceded this crisis to understand the reasons that led to the decision to proceed with the Ariane program and have subsequently guided the program up to the present time.

The first European initiative in the launcher field was taken by the United Kingdom in 1961, which put forward proposals to France, and later Germany, for manufacturing a launcher based on the Blue Streak, roughly similar to the American Atlas, designed and developed up to that time as a first missile stage. The Blue Streak stage was available, following a change of British policy in this domain. This was the basis for the Europa 1 project that was designed to place payloads of 1 metric ton into low orbit.

Unfortunately, the organization set up already contained the seeds of failure because the United Kingdom was supplying a first stage, whose characteristics were frozen, whereas France took on responsibility for the second stage, and Germany for the third stage. Each adopted its own technology and retained an autonomous stage, creating a total absence of anything resembling a system study or an attempt at general optimization. Furthermore, the program was created by cooperation among sovereign states, coordinated by a secretariat under the direction of a diplomat.

By 1966, there was no lack of technical and financial problems. The British threatened to withdraw from the project, which they did two years later. Therefore, the program was obliged to buy the first stage directly from United Kingdom industrial firms. The program was redirected toward geostationary orbit launches

(a satellite placed into geostationary orbit appears to be in a fixed position with respect to Earth), and a solid propellant fourth stage was added. At the same time, the Australian launch range was abandoned in favor of a new launch center to be constructed in French Guiana. The payload specification was then 270 kg in geostationary orbit, and the new program was dubbed Europa 2.

At the same time, geostationary orbit application prospects became clearer, and in 1970, Europe decided to commence studies and predevelopment work on a more powerful launcher (Europa 3) that could place 1500 kg payloads into geostationary transfer orbit (GTO). Unfortunately, the lessons to be learned from the Europa 1 failures were ignored insofar as program organization was concerned. Furthermore, although the first-stage technical configuration remained prudent, based on experience acquired by France from its Diamant program (first unit acquired in November 1965), the second stage was much too ambitious for Europe and involved a high-pressure liquid hydrogen/liquid oxygen topping cycle engine delivering 200 kN of thrust.

The period from 1969 to July 1973 was difficult for the European space community for the following reasons:

- In the technical context, Europa 1 launches F7, F8, and F9, made from the Australian base, all failed. The launcher on flight F11, the first launch Europa 2 made from the new range in French Guiana, exploded 150 seconds after lift-off on 5 November 1971.
- In the political sphere, Germany decided to withdraw from the Europa 2 program in December 1972 (which led to stopping that program in April 1973) and temporarily suspend its participation in the Europa 3 program. The Germans were also considerably attracted by collaborative proposals made by a NASA glowing from the success of the Apollo program. Germany considered that European efforts in the space transportation sector could be limited essentially to technological development, in parallel with major participation in the post-Apollo program. Nevertheless, negotiations between NASA and European representatives on that subject were difficult and engendered frustrations in Europe. At first invited to be involved in developing specific Shuttle hardware, the European contribution progressively narrowed to a science module that would fit into the Shuttle cargo bay.

France adopted a different approach that expressed a triple objective:

- acquisition of absolute control of space applications;
- founding of this control on an autonomous launch capability, with particular reference to geostationary satellites; and
- adoption of these first two objectives by its European partners to assemble sufficient financial capacity and sufficient volume to make production feasible.

However, these French ambitions were momentarily weakened by two launch failures in the Diamant national program in 1971 and 1973, despite the fact that they followed a run of six successful launches.

Finally, the Ariane program may well owe its very existence to the difficult negotiations undertaken with NASA in connection with the launch of the two “Symphonie” experimental telecommunications satellites. The extremely harsh conditions imposed on the German and French negotiators, including an embargo on using these two satellites for any commercial purpose in particular strengthened the determination to achieve the autonomy proposed by France. European agreement was finally achieved in July 1973, following a period of intensive negotiation.

It was decided to embark on three simultaneous programs:

- the “L3S” heavy launch program proposed by France at the end of 1972 (this program was renamed “Ariane” shortly afterward);
- the Spacelab program in cooperation with NASA, backed by Germany; and
- the “Marecs” maritime telecommunications program backed by the United Kingdom.

The principles of setting up the European Space Agency were also defined, and the Agency was officially formed in 1975.

The commitment made by France in connection with this agreement was very considerable and corresponded to more than 60% funding for the program, plus an undertaking to fund excess cost above 120% of the initial figure of MFF 2,060 (about \$ 447 M 1973) up to a maximum of 15% of this figure. In exchange, France obtained agreement that the European Space Agency (ESA) would delegate management of the program to the French Space Agency (CNES). The obligation to ensure a workload return for each participating country in proportion to its contribution was and is indeed a particular aspect of the ESA programs to be emphasized (Fig. 1).

The Early Days of Ariane

The performance objective for the L3S launcher was rather ambitious, based on the conviction that telecommunication satellites mass values would continue to increase and that Europe had to prepare itself for this evolution. Although the view was unanimous that the geostationary orbit was the most promising for application satellites, the figure of 1500 kg had already been the subject of arguments which were to be reopened at regular intervals when performance enhancements were proposed. The conflict was (and still is) between those who predicted a reduction in satellite mass values under the combined effect of electronic circuit miniaturization and the enhanced performance characteristics of satellite onboard propulsion systems and those who predicted a continued increase in mass values under the effect of traffic growth, congestion of the geostationary orbit, and the resultant reduction in in-orbit transponder cost. Finally, the payload mass objective adopted for the Europa 3 launcher was confirmed. It was set at 800 kg for geostationary orbits, corresponding to about 1500 kg in geostationary transfer orbit (about 200 km by 36,000 km). This represented almost twice the performance of the American Delta launcher, which had previously launched most application satellites for the Western world. This

Country	Contribution %	
	Initial	Final
France	63.87	73.55
Germany	20.12	11.47
Belgium	5.00	4.71
United Kingdom	2.47	2.27
Spain	2.00	2.38
Netherlands	2.00	1.64
Italy	1.74	1.58
Switzerland	1.2	0.83
Sweden	1.1	1.05
Denmark	0.5	0.52

Figure 1. National financial contributions. Financial management of the Ariane program had to ensure a fair return. This meant that each country was to receive a volume of business in proportion to its financial contribution. It was also expected that these figures would remain at the same level for the subsequent production phase. Already difficult to apply because it concerns an objective which is both final—including contingencies—and subject to examination when budgets are voted on for the following year, this constraint was further complicated for Ariane 1 by updating rules which varied from one country to another. For example, Belgium, France, and Switzerland made commitments based on a percentage of initial development cost, whereas Germany undertook to make an annual contribution, expressed in its national currency, which would be revised only once in midprogram, according to monetary parities and observed inflation rates. The United Kingdom was involved only through a specific agreement with France. Furthermore, inflation rates at the time were in double digits, and application of the rule of fair return, while fully understandable in principle, was far from easy. The figure compares the financial contributions at the start of the program with the final contributions.

performance was comparable with that of the Atlas-Centaur, which was used at the time only for few heavy satellites. This decision provided a substantial growth potential for European satellites and demonstrated a determination to design the Ariane launcher for an extended lifetime.

Past experience, as mentioned at the beginning of this article, demonstrated the absolute need for imposing three major principles at the very start of the program, namely, a genuine system approach, strong management, and a simple design.

The System Approach. This made it necessary to regard the Ariane launcher (and its ground support facilities) as a single entity, not merely a stack of independently designed stages; to base the development of these stages and other subsystems on the results of studies conducted at the highest level (trajectory, flight mechanics, general loads, thermal, guidance and control, etc.); to

design electrical systems in terms of the complete launcher, installing only actuator devices and equipment specific to its flight phase in each stage; and to impose design rules common to each discipline. For example, it was in this way that the need for POGO control systems was demonstrated at the beginning of the development phase. This meant that corrector devices could be designed and integrated into the propulsion systems at the outset. This was of even greater interest because Ariane was the first launcher designed entirely for geostationary orbit missions, in contrast to the other launchers existing at the time that derived to a greater or lesser degree from ballistic missiles. CNES entrusted a specific contractor (Aérospatiale) with this task and also associated Aérospatiale with the reviews conducted during the development of the various launcher systems and subsystems.

Strong Management. After obtaining delegated management authority, CNES spent the first year of the program establishing the basis for a strong management structure founded on a number of specific principles:

- unique management link between CNES, main contractors and other contractors;
- clearly defined industrial organization, with precise definition of the tasks allocated to each party. For the Ariane 1 program, CNES used the same French level 1 contractors as for its Diamant program, thus ensuring design unity in the main disciplines. These were Aérospatiale for the launcher stages, SEP (Société Européenne de Propulsion) for the propulsion systems, Matra for the vehicle equipment bay, and Air Liquide for the cryogenic third-stage tanks. Other contractors subsequently achieved level 1 status, including Contraves (Switzerland) for the fairing, DASA (Germany) for the second-stage and then the Ariane 4 liquid propellant boosters, Fiat-BPD (Italy) for the Ariane 3 and 4 boosters, etc.;
- a common understanding of the content of work to be done by each contractor, including achievements expected and reports to be submitted.

These principles led CNES to issue a set of management specifications applicable to each launcher system, subsystem and component, and also its ground facilities. These specifications covered overall planning and milestones, work breakdown structure, industrial organization, technical work coordination, schedule and cost reporting, monitoring of the development of critical elements, quality, and reliability. Application of such specifications in the context of a program involving 10 countries that differ in language and culture was a new departure for Europe and generated initial difficulties with companies each of which had their own particular methods of working. Nevertheless, this proved essential to maintain both visibility and coherence. These basic principles were also applied to the Ariane complementary development programs. They proved their effectiveness and made it possible to achieve effective control of technical development, costs, and time schedules.

Simple Design. This final point meant that the Ariane design should be based only on technologies which were either already available or involved only low development risks. The Europa 3 first stage had been designed on this principle,

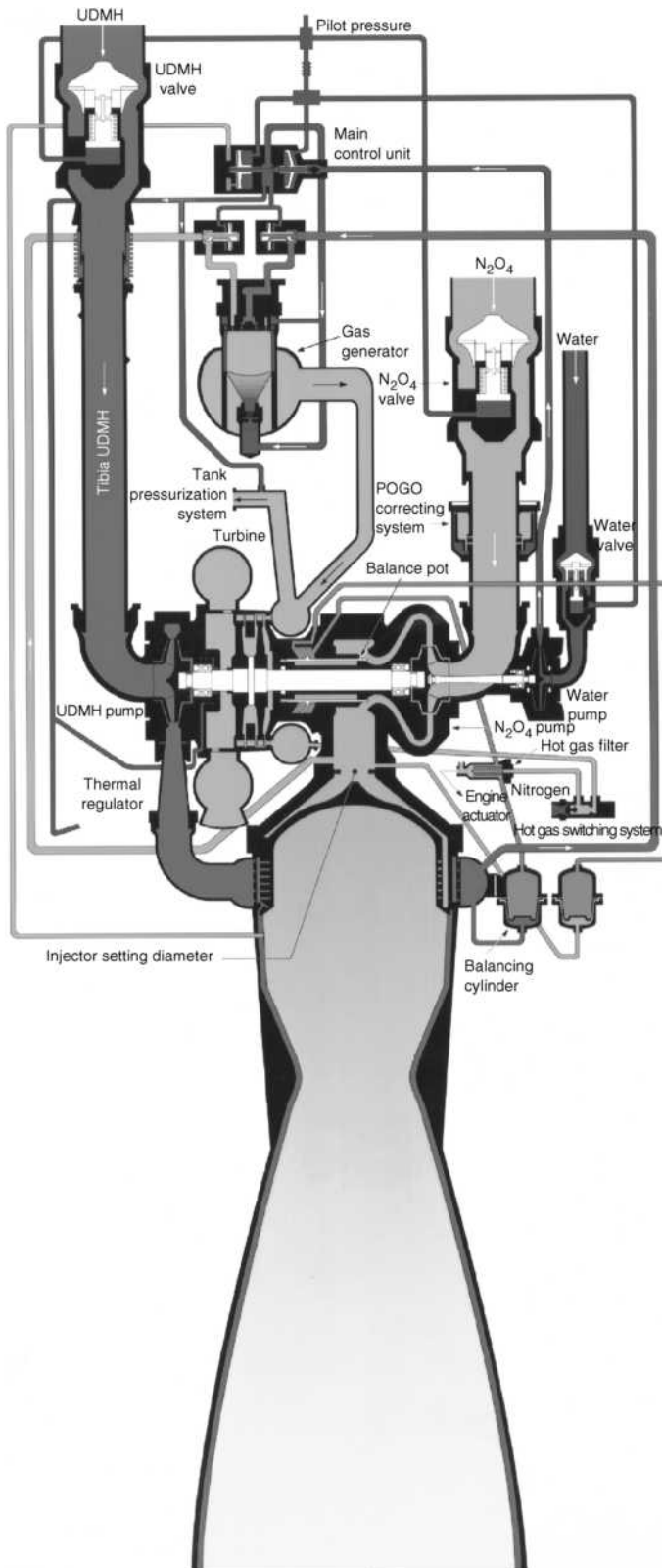
applying technologies and experience acquired with the first stage of the French Diamant launcher. This stage was powered by four rugged “Viking” storable propellant engines, for which the prototype, designed in France, had been tested with very encouraging results before this program was terminated. Therefore, the Europa 3 stage was selected as the basis for the first stage of the Ariane launcher.

This could not be the case for the Europa 3 second stage that was considered much too ambitious for Europe in 1973. However, although a number of difficulties were to be anticipated if liquid hydrogen and liquid oxygen were used, the efficiency of these propellants would make it possible to reduce the lift-off mass of the launcher by a factor of almost 2. Furthermore, a certain amount of experience had been acquired on a low-thrust gas generator cycle prototype engine developed within the French national program. Work conducted in France and Germany during the Europa 3 program meant that this approach could be adopted with a satisfactory level of confidence, provided that there was no major deviation from the experience acquired from the engine and propellant and fluid system. Thus, the tanks were designed for 8 tons of propellant and a maximum possible diameter using available tooling not to exceed 2.6 m. Having made this choice, it was found necessary to add a second, intermediate stage to achieve the performance target. This stage was designed on the basis of elements and technologies developed for the first or third stage, including the Viking engine used in the first stage (with a suitably adapted nozzle) and the light alloy tanks of the third stage.

Ariane 1 Launcher

Ariane 1 was a three-stage launcher that had a total height of 47.8 m and a lift-off mass of 210 tons. The first stage had a dry mass of 13.3 tons, a height of 18.4 m and a diameter of 3.8 m. It had four Viking 5 engines (Fig. 2) that developed 2500 kN thrust on lift-off. Burn-time in flight was 146 s. The 148 tons of propellant (UDMH and N_2O_4) were contained in two identical steel tanks protected against internal corrosion by an aluminum layer and connected via a cylindrical skirt. The four turbopump Viking engines were mounted symmetrically on a

Figure 2. Viking engine flow diagram. The Viking engine uses the gas generator cycle, and the gas generator itself operates at a stoichiometric ratio. The gas produced is cooled by injecting water to reduce gas temperature to values compatible with the turbopump turbine, pressurization of the main propellant tanks, and operation of the power pack for the hydraulic servoactuators. The three pumps (UDMH, N_2O_4 , and water) are mounted on a single shaft that rotates at 10,000 rpm. (For Ariane 1, the chamber pressure was limited to 53.5 bars.) The hydropneumatic regulation system slaves the combustion pressure to a reference pressure value by adjusting the gas generator feed and, thus, the rotational speed of the turbopump. The mixture ratio is maintained at a constant level by a regulator that equalizes propellant pressures before injection into the combustion chamber. Engine ignition is induced by the pressure in the propellant tanks. When the valves open, the propellants, which are hypergolic, are delivered to the combustion chamber and gas generator and ignite spontaneously. The turbopump speed then builds up to the value set by the regulating system in 1.3 s. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



thrust frame and articulated in pairs on two orthogonal axes to provide for three-axis attitude control. An annular water tank, located inside the propulsion bay, provided for cooling the gas obtained from the engine gas generators that was used to pressurize the propellant tanks and supply the hydraulic motors of the attitude control actuator systems. Four fins with a surface area of 2 m^2 provided aerodynamic stability. This first stage was designed to destruct approximately 30 seconds after stage one to two separation.

The second stage had a dry mass of 3.13 tons (excluding the interstage conical skirt and the jettisonable acceleration rockets), its height was 11.6 m, and its diameter was 2.6 m. It had a single Viking 4 engine that developed 740 kN thrust in vacuum for a burn-time of 136 s. The motor was linked to the conical thrust frame via a gimbal with two degrees of freedom, that provided pitch and yaw control. Auxiliary nozzles supplied with hot gas from the engine gas generator provided the roll control function. The two aluminum alloy tanks that had a common intermediate bulkhead were pressurized with helium gas (3.5 bars) and contained 34.1 tons of propellant (UDMH and N_2O_4). The second stage was also designed to destruct about 30 seconds after stage two to three separation. During the prelaunch waiting period on the pad, a thermal shroud, ventilated with cold air, which restricted heat exchange between the propellants and the environment, protected the second-stage tanks. This shroud was jettisoned on launcher lift-off.

The third stage weighed 1.164 tons dry, was 9.08 m high, and had a diameter of 2.6 m. This was the first cryogenic stage produced in Europe. It was equipped with a type HM7A engine that developed 62 kN thrust in vacuum for a burn-time of 545 s. This engine was designed by Société Européenne de Propulsion (SEP), based on experience acquired with an earlier cryogenic engine, the HM 4, which delivered 40 kN thrust, and tested over the period 1962–1969. The HM 7a engine uses the conventional gas generator cycle technology, and achieves a specific impulse of 443 s with a mixture ratio of 4.43 at pump inlet. The combustion chamber is supplied with propellants pressurized by a turbopump, via a set of injection valves. The pump turbine is driven by gas supplied by a generator, the latter receiving a small proportion of propellant tapped off at the pump outlet. The liquid hydrogen and liquid oxygen tanks that contained a total of 8.23 tons of propellant were made of an aluminum alloy and had a common intermediate bulkhead (double skin under vacuum). The tanks were covered with external thermal protection to avoid warming the propellant. Both tanks were pressurized in flight, using hydrogen gas tapped at the outlet from the regenerative chamber and helium. The motor was linked to the conical thrust frame via a gimbal that provided pitch and yaw control. Auxiliary nozzles ejecting hydrogen gas were used for roll control.

The stages were separated by pyrotechnic cutter devices fitted on the rear skirts of the second and third stages. The separating stages were distanced from each other by retrorockets incorporated in the lower stage and acceleration rockets mounted on the upper stage. Stage 1 to stage 2 separation was controlled by the onboard computer, on detection of first-stage thrust decay (propellant exhaustion). Stage 2 to stage 3 separation was controlled by the onboard computer when the second-stage speed increase reached 1500 m/s.

The vehicle equipment bay (VEB) weighed 316 kg, had a diameter of 2.6 m, and was of 1.15 m high. Mounted on the third stage, the VEB contained the

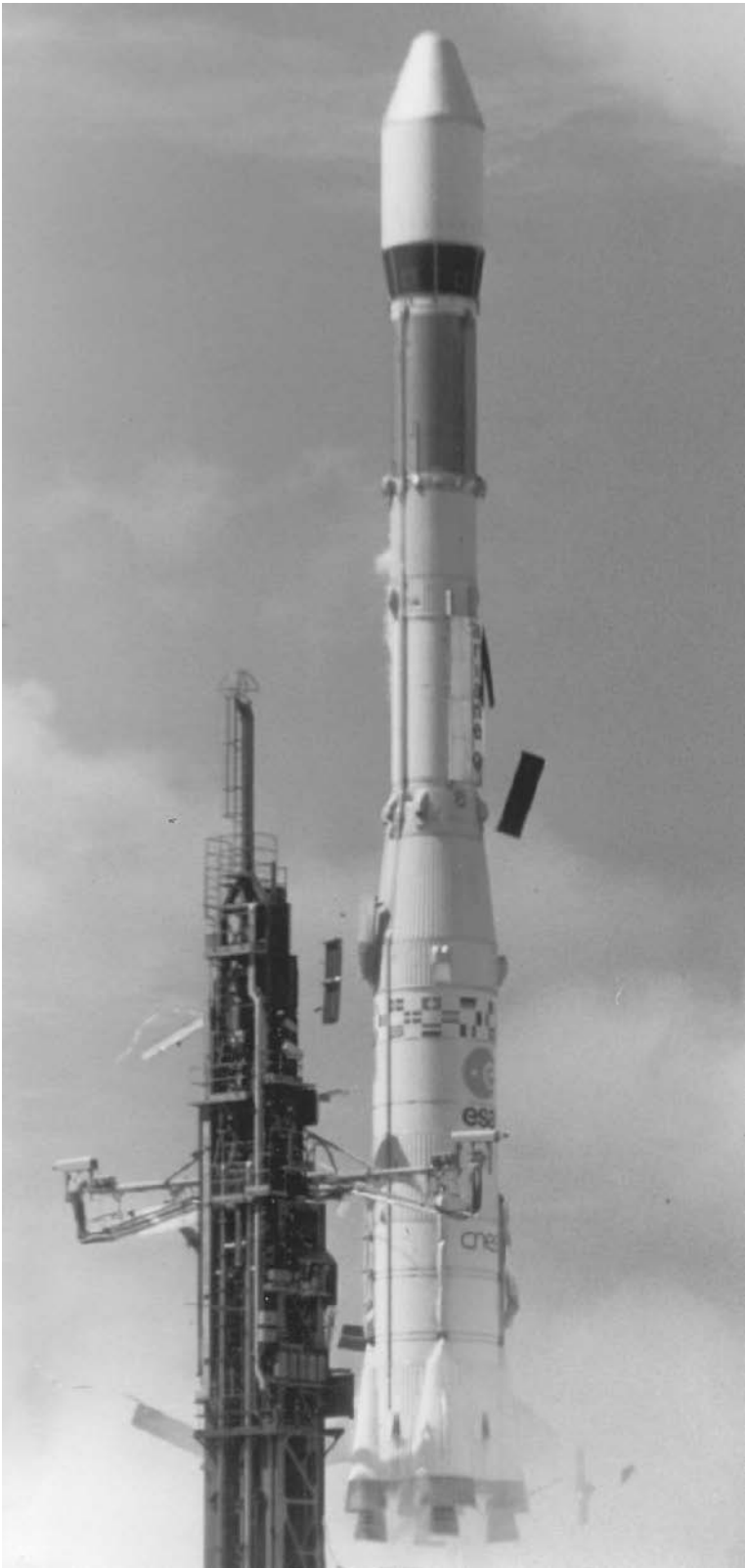
electronic equipment of the launcher and also served as a base for the payload and fairing. In addition to the onboard computer, the VEB housed all of the electrical equipment required for executing the launcher mission, namely, the sequencing unit, guidance and navigation control, and the location, safety, and telemetry systems. Only the power systems and actuator devices were located elsewhere in the launcher stages.

The two half-fairings were ejected parallel to the main axis of the launcher under the control of the onboard computer, as soon as the calculated thermal flux dropped below the specified level. The fairing was jettisoned by two pyrotechnic systems, a horizontal system at the interface with the VEB and a vertical system which also served to impart horizontal velocity to each half-fairing. The fairing was 3 m in diameter and was compatible with Atlas-Centaur class satellites.

The launch facilities in French Guiana (ELA 1: Ariane launch site No1) were designed to make use of the earlier investment for the Europa 2 launcher. The stages were erected and assembled directly on the launch pad. A mobile gantry sheltered the launcher and provided for assembly of the payload with the launcher and closure of the fairing under clean room conditions. The gantry withdrawal commenced 6 hours before lift-off. The main fueling of the third stage with liquid hydrogen and liquid oxygen and final topping off were performed using a cryogenic arm system carrying a set of umbilical valve plates. Disconnection and retraction of the arms commenced at time $T - 3$ s. Control of the launcher on the pad was fully automatic from time $T - 6$ minutes. The launcher was then controlled by two ground computers, one for the electrical systems and the other for the propellant and fluid systems. The two computers also cross-checked each other. The first stage engine was ignited at time T , and the lift-off command signal was sent at $T + 3$ s, following satisfactory verification of the Viking engines. The first flight took place on Christmas eve 1979 (Fig. 3).

When the Ariane program was first initiated in 1973, few imagined a commercial career for the launcher. Things began to move in 1976 and led to a series of actions relating to launcher performance, production, and marketing. As regards performance, the idea was to propose Ariane for launching the Intelsat satellites. The success of such a venture would represent both an exceptional reference for the program and strong motivation for the players involved. However, the first task was to augment performance objectives up to 1600 kg minimum in GTO. Fortunately the prudent approach adopted for the basic definition of the launcher, combined with the results of the first flight, demonstrated propulsion performance in excess of specifications and made it possible to exceed the initial objective and achieve a figure of 1850 kg.

As far as production was concerned, it was obvious that there was no hope of selling the launcher if one waited to receive orders before commencing manufacture. It was on these lines that discussions were opened at the European Space Agency, and a promotional phase was duly adopted. Apart from the production of six launchers, the objective of this phase was to achieve full operational qualification, develop and validate the dual launch capability, and adapt the launcher comprehensively to meet user needs by providing for the construction of payload preparation facilities in French Guiana. Analysis of marketing aspects led to two actions: initiation of the Ariane 3 program and formation of Arianespace.



Ariane 3 Program

At the end of the 1970s, the application satellite market was organized around two launcher groups. These were the Delta launchers, for which GTO performance in 1978 was around 1200 kg and the Atlas-Centaur launchers whose performance figure was close to 1800 kg. Apart from the Intelsat satellites, all other payloads corresponded to the Delta class, for which Ariane could not pretend to be competitive in its actual state. The idea then emerged to adapt the launcher to enable it to launch two Delta class satellites simultaneously, thus raising the performance objective to twice 1200 kg, plus the mass of the dual launch structure designed to isolate the two satellites from each other. This corresponded to a total GTO mass of 2500 kg. The launcher also had to be competitive with the Atlas-Centaur.

This program was proposed by France to the European Space Agency and was approved in July 1980, despite the failure of the second Ariane flight (L02) in May of that year. CNES was charged with managing the Ariane 3 program, under conditions very similar to those for Ariane 1.

Analysis conducted from 1976 and aimed at increasing the performance of Ariane 1, had in fact identified modifications whose feasibility was checked out during the final development test phase. These modifications were adopted in part or in full for the Ariane 3 program, in accordance with the policy of minimizing development risks. The addition of two solid propellant boosters achieved

Figure 3. Launch L0 1 (24 December 1979). The decision to proceed directly to flight tests, using three active stages in the final flight configuration, was made following very lengthy discussions at the start of the Ariane program. This was in total opposition to the highly progressive, but also extremely costly approach adopted for the Europa 2 program. Initial ignition occurred on 15 December 1979, following a faultless countdown. Unfortunately, the ground computers did not authorize liftoff, and the engines shut down automatically at time $T + 10$ s and aborted the launch. Subsequent analysis showed that an explosion in a small measurement pipe had damaged the sensors, whose signals were used for operational diagnosis of the engines. The case of an aborted launch, for which the probability was infinitely small, had, nevertheless, been taken into account during the development phase. Procedures for return to flight configuration had been written and validated by tests conducted in Europe. This allowed restarting the count on 23 December, following 8 days of round-the-clock work. Technical problems, combined with adverse meteorological conditions, prevented a launch on 23 December. On the following day, Ariane made a practically faultless launch. Only two anomalies were identified, and these were corrected before the following flight. These concerned minor pollution of the payload, caused by the second-stage retrorockets and low amplitude vibrations (POGO effect) at the end of the second-stage flight. The first Ariane launch thus took place only 6 months after the initial target date set in 1973.

The second launch in May 1980 resulted in a failure. Destruction of the launcher occurred at $T + 63.75$ s, as a result of high frequency combustion instability in one of the first-stage Viking engines, 2.75 seconds after liftoff. Corrective measures essentially comprised modifying the propellant injection orifices in the combustion chamber. A total of 95 tests that represented 4300 seconds of burn-time were conducted to qualify the new injector under conditions substantially more severe than those encountered in actual flight. The Ariane 1 development phase terminated in December 1981, following three successes out of the four launches made. The total cost of the program was within 120% of the initial estimate. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the objective of competitiveness, with Ariane 3 for two satellites of 1200 kg each, and Ariane 2, with no boosters, for satellites of 2000 kg.

These performance objectives were achieved on the basis of the Ariane 1 launcher by introducing the following modifications:

- increased thrust for the first- and second-stage Viking engines by augmenting combustion pressure by 10% (53.5 bar to 58.5 bar). This was obtained mainly by adding hydrazine hydrate to UDMH;
- adding two solid propellant boosters with a unit thrust of 600 kN and a burn-time of about 40 seconds. The low level of performance sensitivity to the structural mass of the boosters made it possible to adopt extremely prudent technical solutions;
- increase in the third-stage propellant load from 8 to 10 tons;
- enhancement of Stage 3 performance by increasing combustion chamber pressure by 5 bar and stretching the nozzle by 200 mm (HM7 B);
- adaptation of the SYLDA dual-launch system and the fairing to the volume required for 1200 kg-class-satellites.

Figure 4 shows a SYLDA dual-launch structure and half a fairing. The lower satellite is placed inside an egg-shaped compartment. Protected by the fairing, this carbon fiber structure is subject to reduced loads by comparison with an external structure. Furthermore, the operational constraints induced by the need to carry two satellites remain within acceptable limits.

The SYLDA structure makes it possible to achieve complete separation of the two payloads. The long orientation sequence, spin-up phase and separation of the satellites and upper part of the SYLDA, are achieved by the SCAR attitude and roll control system, using third stage pressurization hydrogen gas to operate this system.

Tested successfully on the sixth launch, this concept was unquestionably one of the keys to the success of Ariane. The cost of a launcher is not proportional to its size, and a number of functions must exist whether the launcher is small or large. A dual-launch capability, taking advantage of this scale effect, represents a major competitive plus factor.

The industrial organization was the same as that for the Ariane 1 program; the only main modification was an increase in the Italian contribution, which led to entrusting the development of the solid propellant boosters to FIAT-BPD. The first Ariane 3 was launched successfully in mid-1984.

Arianespace

The lengthy discussions which took place at the European Space Agency and led to the decision to go ahead with production of six Ariane 1 launchers, also demonstrated that this multinational organization, whose basic purpose was research and development, was not suitable for engaging in commercial and production activities. At the same time, the promoters of Ariane reached another paramount conclusion. For European autonomy to be effective in access to the geostationary orbit, the Ariane launcher had to be credible. To this end, it had to be both reliable



Figure 4. SYLDA, Ariane dual-launch system. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

and available, in other words produced in sufficient quantities and in advance of actual launch needs. This called for world-scale marketing, and at the same time made it possible to spread the fixed production costs over a larger production volume, and thus reduce the cost of the autonomy-related strategy (in contrast to its American competitors, Ariane did not have a major captive market for launching governmental satellites, to which all or part of the fixed costs could be allocated). Only a private commercial entity, responsible for both Ariane production, marketing, and launch operations, could take up this challenge.

In December 1977, Frédéric d'Allest, then head of the CNES launcher division and future Chairman and CEO of Arianespace, proposed forming a company to market the Ariane launcher. *Arianespace* was incorporated on 26 March 1980, well before qualification of the Ariane 1 program. *Arianespace* is a French business corporation. Its capital is held by CNES, the leading European manufacturers that participate in Ariane production, and a number of banks. *Arianespace* is responsible for producing, launching, and marketing the Ariane launcher or launchers, whose development and qualification have been, are, or will be conducted by the European Space Agency.

The first Arianespace commercial launch was made successfully on 22 May 1984, inaugurating the very first commercial space transportation service.

Quality

A number of technical difficulties were encountered during the early days of the Ariane program, resulting from residual design problems or omissions in the production files. The European launcher industry was in its infancy was in the process of discovering the problems inherent in launcher batch production, and had to acquire expertise in this demanding discipline. Major efforts were made during the 1980s to amplify and correct the production files, identify and analyze all defects irrespective of their importance, and carry out regular tests on equipment sampled from the production line. The severity of these tests frequently exceeded the levels specified for the qualification tests. The flight data measurement plan that covered more than 700 parameters for which results were transmitted to the ground during the qualification flights—this figure was reduced to 400 for operational flights—made it possible to analyze each flight in the finest detail and thus acquire in-depth knowledge of the launcher. This unprecedented quality construction program, conducted by a European team highly motivated at all levels, meant that the 90% reliability objective assigned to the Ariane in 1973 was quickly exceeded. By the end of October 1999, a total of 118 Ariane launchers (versions 1 to 4 inclusive) had flown, and the reliability figure for Ariane 4 exceeded 97% after a string of 48 consecutive successful flights.

Marketing

Ariane marketing operations achieved rapid success, despite a number of technical difficulties encountered at the start of the program. The success of the development plan, combined with the appraisals conducted by potential customers, bred a high degree of confidence in the European launcher and the French Guiana ground facilities. The first non-European customer to place an order for Ariane launch services in 1978 was Intelsat (International Telecommunications Satellite organization), well known in the satellite world for its technical expertise. This choice was extremely important for Ariane and induced confidence on the part of other customers, who then decided to regard the Ariane launch system objectively. Apart from its own inherent merits, Ariane had the advantage of a favorable market situation. Commissioning of the American Space Shuttle led NASA to stop producing its conventional Delta and Atlas-Centaur launchers. Delays with the Shuttle, due to technical difficulties, discouraged satellite operators, obliging them to seek other launch possibilities. Furthermore, the high U.S. dollar exchange rate during the early 1980s made Ariane prices extremely attractive compared with the American launch systems. This marketing success grew further with the passage of time. The first Ariane 3 flight in August 1984 qualified the dual-launch capability of two Delta class spacecraft and significantly increased both the launch capacity and competitiveness of the Ariane system.

In 1985, Ariane launched the same number of commercial satellites as the Space Shuttle, and in the following year, *Arianespace* signed no fewer than 16

launch contracts. By 1987, Arianespace had 57 contracts (satellites to be launched or already launched) from the European Space Agency, the member countries of the European space community, international organizations, national organizations in non-European countries and private companies, representing a total sales value exceeding FF 16 billion (about \$2.25 billion). *Arianespace* had acquired a share of more than 50% of the open satellite market, a position that the company has succeeded in maintaining despite increasingly severe competition.

Ariane Launch Site No. 2 (ELA 2)

France decided to construct a launch range in French Guiana in April 1964, after the deciding to discontinue launch operations from the Hammaguir range in southern Algeria. The French Guiana site was chosen as a result of a comparative study of a number of possible locations. Paradoxically, if we compare the 1964 situation with that of today, equatorial launches, if they were even mentioned, did not initially constitute a priority criterion. However, this view was quick to change, and in July 1966, the European launcher organization accepted the French proposal to contribute to constructing an equatorial launch base in French Guiana, initially intended for Europa 2. It should be noted that at that time, 33 years ahead of the Sea Launch project, the French Guiana site was in competition with a floating marine platform project proposed by Italy, similar to the San Marco platform which it was then operating offshore from Kenya.

A decision was quickly made in favor of the Kourou site (the population of the village of Kourou was 660 in 1964) on the Atlantic seaboard 70 km northwest of Cayenne in a sparsely populated region. The exceptional geographical characteristics of this site combined a wide launch arc over the ocean and favorable climatic conditions (infrequent storms, no cyclonic or seismic activity, and temperatures varying only slightly round a mean figure of 25°C). All types of mission (polar and equatorial orbits in the range -10.5° to $+93.5^\circ$) could be planned in complete safety, an advantage that no other operational base possessed then. The closeness of Kourou to the equator ($5^\circ 2' \text{N}$) is ideal for placing telecommunications satellites into geostationary orbit. At this latitude, the slingshot effect induced by the rotation of Earth is near its maximum, and propellant consumption for adjusting the plane of the geostationary orbit is minimum. In global terms, the mass gain achieved with a launch from Kourou is approximately 17% compared with a launch from the Cape Canaveral, using an identical launcher.

The Guiana Space Centre (CSG) facilities were first operationally tested in April 1968 with the launch of a Véronique sounding rocket, and the Centre was inaugurated officially in 1969. The first satellite was launched on 10 March 1970 with the French Diamant B launcher. The first Europa 2 flight was in November 1971, using the new CSG launch pad. The failure of this launch had a series of consequences already mentioned at the beginning of this article. For both economic and political reasons, it was essential for the Ariane 1 launcher to make the best possible use of the heavy investments made in Europa 2. Nevertheless, certain facilities were unsuitable for launch operations with Ariane because of insufficient capacity and lack of operational flexibility and due to an absence of

any growth potential. This made these facilities largely incompatible with the Ariane performance enhancements already under study. After proposals from the French Space Agency (CNES), the European Space Agency (ESA) decided to construct Ariane launch site No. 2 (ELA 2) in July 1980. This new facility was required to meet the following specifications:

- provision for 10 launches per year (Ariane 2, 3, or 4) and execution of two launches within an interval of less than one month;
- provision for replacing a launcher already on the pad, in case of need, by the launcher scheduled for the next flight;
- provision for preparing larger payloads with improved facilities for payload/launcher integration.

The design of the new launch site differed from that of ELA 1. The time spent by the launcher on the pad had to be reduced to a minimum to allow for executing launch operations and postlaunch rehabilitation of the pad within a maximum 1 month. This assumed locating a launcher preparation zone on the edge of the safety perimeter.

Following their arrival from Europe, the stages are erected, assembled, and tested in the rear preparation zone. Then, the launcher on its mobile launch table is transferred to the pad via a dual rail track (Fig. 5). There it is connected to the propellant and fuel circuits and undergoes a launch rehearsal procedure, including filling of the third stage with liquid hydrogen and liquid oxygen to check the absence of ground and onboard leakage. The success of this operation authorizes assembly of the upper part (payloads, adaptors, and fairing were integrated beforehand in a dedicated building) with the launcher. The principles of



Figure 5. ELA 1 and ELA 2 launch sites. ELA 1 site (foreground): Ariane 3 launcher during third-stage fueling tests on day D – 9. At rear: Ariane 4 launcher recently arrived on the ELA 2 pad. Note the double rail track between the pad and the preparation zone (background). The white circle midway down the rail track is a turntable, used to allow two launchers to pass each other if necessary. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the launch sequence qualified on the ELA 1 site have been retained. The ELA 2 pad uses many of the support facilities previously used for ELA 1.

Ariane 4 Program

Ariane 3 represented a short-term response that enabled the Ariane launcher to position rapidly vis à vis its two American competitors, Thor Delta and Atlas-Centaur. The arrival of the Space Shuttle at the end of the 70s and the policy adopted by NASA of marketing Shuttle flights at extremely attractive prices required a more comprehensive response from the Ariane program.

It was obvious that the NASA price policy would lead to a profound change in satellite design, broadly on the following lines:

- increase in Delta class satellite size and mass, up to the point of occupying, in a vertical position, the maximum volume available in the Shuttle cargo bay. This led to satellites of between 1400 and 1500 kg for injection into geostationary transfer orbit;
- appearance of a family of satellites installed in a longitudinal position in the Shuttle cargo bay, designed to take advantage of the large diameter available and reduce launch cost (dependent on the height of the satellite). Corresponding changes, that involved problems of both mass and diameter were more difficult to predict.

As regards satellite mass, there were some projects for TV satellites of 2500 kg. However, these were rare and still only moderately credible at the end of 1980, in particular among those who predicted a reduction in satellite mass values from miniaturization of satellite electronics. However, things moved in 1981, and the objective of 2500 kg was finally adopted.

The diameter problem was even more delicate. The Ariane program team foresaw major aerodynamic problems if the fairing diameter exceeded by too great an extent the 2.6 m diameter of the third stage on which it was mounted. However, determination of this value required aerodynamic tests that were too lengthy to conduct before making the decision. The diameter of 3.65 m finally available for payloads was in fact the result of a compromise between what the program team considered possible at the least risk and the sacrifice which satellite customers were prepared to accept for the advantage of a second launch service source.

Competitiveness requirements dictated the pursuit of the dual-launch policy implemented for Ariane 3. The GTO performance objective assigned to Ariane 4 was consequently an ability to execute simultaneous orbit injection of two payloads, one of 1400 kg and the other of 2500 kg, giving a total of 4300 kg, including the mass of the dual-launch structure. (Note that this objective had increased from 3400 to 4300 kg in 1981 and that the maximum performance demonstrated in 1999 exceeded 4900 kg.)

This program was to be regarded as a continuation of the Ariane family, in other words as a complementary performance enhancement phase, taking full advantage of work carried out for Ariane 2 and 3, and innovating as little as possible in propulsion. This was decided to take the fullest advantage of

experience already acquired and to improve the reliability target from 0.90 to 0.92. Given the ambitious performance objective that had been set, the configuration adopted was required to embody a certain degree of flexibility with regard to lower mass payloads to apply an attractive pricing structure for a wide range of satellites. Single launches of heavier satellites like Intelsat 6—3600 kg—had to be possible, of course.

These performance objectives were achieved on the basis of Ariane 3 by introducing the following modifications:

- First stage propellant load increased from 147 to 227 tons, while retaining the same operating point for the Viking engines as qualified for the Ariane 3 program. Stage 1 burn-time was increased to 205 s.
- Development of a liquid propellant booster, corresponding to a reduced scale copy of the first stage as it uses the same engine with an appropriately adapted pressure ratio, and the same tank pressurization system. Identical stainless steel tanks carry 39 tons of propellant.
- Adaptation of the Ariane 3 solid propellant boosters. Carrying 9.5 tons of propellant, these boosters burn for 36 s and deliver 650 kN thrust each.
- Development of a new water tank, located on top of the UDMH tank. This tank is constructed in composite materials, and supplies the first stage and liquid propellant boosters in blow-down mode.
- Modification of the vehicle equipment bay structure, to achieve better payload integration flexibility and easier transition to the new fairing diameter.
- Development of an external Ariane dual-launch carrier structure (SPEL-DA), providing for dual launches of large-diameter satellites.
- Development of a new fairing with a useful diameter of 3.65 m.

This program was proposed by France to the European Space Agency and formally accepted in January 1982. The same program management principles were adopted as those for Ariane 1 and 3.

Launcher Development. The development phase did not introduce any major innovations in propulsion. First, the first-stage propulsion bay had performed extremely satisfactorily on the test bed in the Ariane 3 development program for burn-times very close to those required for Ariane 4. Furthermore, the booster propulsion system was based on a configuration already used to develop the Viking engine. Consequently, insofar as propulsion was concerned, the task was one of optimization or of demonstrating new operating margins because the burn-time for the Viking engine had been increased substantially. On the other hand, work relating to the launcher system was substantially more complicated than first thought. It was found necessary to adopt the complete system approach for each lower composite configuration. Furthermore, the increase in the height of the launcher and the size of the upper composite were reflected in a considerable increase in general loads that required substantial modification of a number of structures, including connecting flanges, in particular. Furthermore, first-stage in-flight stability was a major source of concern during the first few years of the program. Digital control was adopted to introduce a large degree of flexibility in the development time schedule and also with actual operation of the

launcher by making it possible to finalize the configuration of each launcher only 2 months before lift-off. Today, Ariane 4 can place from 2.1 to 4.9 tons in a typical GTO orbit, depending on the number of liquid (L) or solid (P) propellant boosters fitted on the first stage (AR 40,42P,42L,44P,44LP,44L). Nine different configurations of SYLDA, SPELDA, and fairings were available for payload accommodation. The Ariane 4 maiden flight, initially planned for late 1985, took place successfully on 15 June 1988.

The Birth of Ariane 5

Initial reflection at CNES on Ariane 5 dates back to 1978. At that time, the Ariane 5 launcher was regarded more as a means to access low-orbit, manned-flight missions. Nevertheless, the launching of application satellites into geostationary transfer orbit continued as an objective, in particular in triple-launch mode, to pursue the Ariane scale effect competitiveness approach. The two missions rapidly acquired identical importance.

In contrast to the approach adopted for the American Space Shuttle, the primary principle adopted was that priority should be given to competitiveness for commercial flights. Manned spaceflight induces substantial costs that it is absurd to impose on operations that can be conducted by an unmanned vehicle. The system proposed had to be capable of manned flight and unmanned flight missions without increasing the cost of the latter. Consequently, the new launcher had to be regarded and optimized as a single entity at the design stage to ensure optimum use of available resources, and the manned flight aspect had to benefit from lessons learnt from unmanned missions, while also providing for substantially dissociated use when the operational phase was reached. The manned module was consequently designed as a “special payload” mounted on top of the launcher. The capsule and spaceplane concepts were then analyzed in parallel. The latter presented the advantage of a considerable reentry cross-range and a high degree of orbit return flexibility that provided for a soft landing and consequent reuse. The onboard intelligence provided launcher “brain” functions during the ascent phase and thus eliminated the need for a vehicle equipment bay required for unmanned flight. This, then, was the Ariane 5 - Hermes concept at the end of 1979.

At that time, Ariane 5 comprised the Ariane 4 first stage, and a new second stage burning liquid hydrogen and liquid oxygen (H 55), carrying Hermes for manned missions or the Ariane 4 third stage (H 10), the vehicle equipment bay, and fairing for unmanned flight. This study, which emphasized the need for a large cryogenic engine for future Ariane improvements, made it possible to propose and obtain funding for a 3-year French national program in 1980 and to commence work on the basis of design for an engine to deliver 600 kN thrust. Germany, Sweden, and Belgium joined this project under the terms of bilateral agreements.

Partially Reusable Concept Studies. In 1980, the impact of the Space Shuttle and its price policy led to strong criticism of the conservative characteristics of the Ariane 5 configuration. As a matter of principle, reusable concepts were regarded as more economical, the more so because the maintenance costs for such systems were ignored and operating costs were substantially underestimated. Considerable attention was then paid during that year to improving the

cost levels and analyzing the operating costs announced for the Space Shuttle and those observed for the early stages of the Ariane operational phase. In parallel, the problems induced by the rehabilitation of equipment used during the Ariane 1 development phase demonstrated the importance of the corresponding work and the complementary qualification cost that would be required.

A number of concepts were considered:

- recovery of the first stage by parachutes. This was tried unsuccessfully with the first stage of the Ariane launcher used for flight 14. The liquid propellant boosters for Ariane 4 were also initially designed to facilitate this type of recovery.
- design of a first stage comprising a stack of several H 55 stages, recovered individually using a delta wing;
- consideration of a winged first stage, etc.

This line of approach failed to produce any attractive, realistic solutions, and was abandoned early in 1982.

Comparative Configuration Studies. A systematic review of all possible concepts that matched the performance objectives was then undertaken. More than 24 different configurations were drawn and assessed for performance and cost. A configuration involving a large cryogenic core stage flanked by two large solid propellant boosters was examined and then abandoned because it was impossible to identify a derivative solution, with reduced performance, which would have made it possible to adapt a launcher configuration to the mission model in the same way as with Ariane 4. This configuration was looked at again, when the desired degree of flexibility had been introduced by offering the choice of an Ariane 4 third stage (H 10), or alternatively a highly simplified storable propellant stage, and after further studies had made it possible to set an acceptable price objective for manufacturing the solid propellant.

Finally, these three configurations were selected for more detailed comparative analysis in mid-1983:

- a solution that corresponded to direct continuity with Ariane 4 but introduced a cryogenic second stage. This solution was derived directly from experience acquired from Ariane 4 and made it possible to evaluate cost and performance extremely precisely.
- a solution based on the above, replacing the Ariane 4 first stage with a cryogenic stage equipped with four or five engines identical to that used for the second stage. This solution had the advantage of requiring only development of a single propulsion engine, namely a high-thrust LOX/LH₂ engine.
- a solution involving a large cryogenic core stage, flanked by two large solid propellant boosters. This basic composite carried a simple storable propellant upper stage for low-performance missions or a cryogenic stage that carried 10 tons of LOX/LH₂ (Ariane 4 third stage) for high-performance missions.

All three configurations required developing of a new LOX/LH₂ engine. This conclusion led France to propose the development of an engine of this type, based

on the project commenced in 1980, to the European Space Agency. A further year was then devoted to a detailed comparative study of the three configurations.

The configuration that had large solid propellant boosters was finally selected on the basis of intrinsic reliability, recurrent cost, greater potential, and development time schedule control criteria. Only the safety criterion remained questionable. Compared with the abrupt failure of solid propellant boosters, their liquid propellant counterparts are always presented as easy to control. Nevertheless, in-depth analysis of the complete system demonstrates that this is not always the case, and that even in this favorable situation, there are still flight phases where the spaceplane (in contrast to a capsule) simply cannot accomplish its true rescue mission. (Long after this choice had been made, this frequently impassioned discussion was stimulated once more following the Challenger failure.)

Therefore, the configuration that incorporated two solid propellant boosters was finally presented in 1985 and adopted as a European Space Agency program in 1986 and 1987. Further optimization led to incorporating the following aspects:

- abandonment of the triple-launch concept in a basic mission context due to the excessive operational constraints relating to the satellites (availability);
- increase in payload mass values: the nominal performance objective was then dual launches into geostationary transfer orbit for satellites of 2950 kg each;
- mass problems encountered with the Hermes program. This led to an increase in the performance of the lower composite and abandonment of the cryogenic upper stage.

Ariane 5 Specifications. Final specifications for the Ariane 5 program were as follows:

- simultaneous launch of two satellites of 2950 kg and a diameter of 4.53 m into geostationary transfer orbit under environmental conditions and with a degree of precision, etc., comparable with Ariane 4 launches, representing a single-launch performance equivalent of 6800 kg;
- launch of an 18-ton payload into a circular orbit of 550 km, inclined at $28^{\circ}5'$;
- Hermes launch: this mission was not to introduce constraints liable to penalize unmanned flight. Consequently, performance requirements for this mission were deduced from those specified for unmanned flight, taking due account of safety constraints;
- Reliability of unmanned flight was set at 0.98, almost 10 times higher than the initial specification for Ariane 1. This ambitious target was justified by the high cost of insurance for both launcher and satellites (over 20% in 1988) and the major consequences of a flight failure. The safety of the crew was consequently provided for by ejection of the Hermes spaceplane on detection of an operating anomaly. The safety objective for the crew was set at a figure of $1 - 10^{-3}$.
- Cost objective per launch, on the basis of eight launches per year, was set at 90% of the figure for an Ariane 44 L. Given the respective performance of the two launchers, this corresponded to a reduction of 45% in the cost per kilogram in orbit for an equivalent fill factor.

- The target for a maiden flight was initially set for April 1995. It was subsequently put back by one year due to a number of economic constraints.

As for earlier Ariane programs, management was delegated to the French Space Agency (CNES) by the European Space Agency. The conditions for this delegation of authority were more restrictive in this case, however, due to the need for close coordination with the Hermes program and the now powerful image of the Ariane program.

The three main management principles stated at the beginning of this article with regard to Ariane 1 were retained and in fact strengthened in two areas:

- Management specifications incorporated a design-to-cost objective.
- The safety/reliability approach was further emphasized through systematic integration of past experience, both from incidents and accidents that occurred during testing, in the course of previous Ariane flights, or in the context of other programs (Challenger, etc.).

Furthermore, faced with an identified failure mode, the principle of a dual approach, involving simultaneous reduction of the probability of that failure and improvement of system tolerance was imposed.

Ariane 5 Launcher

The Ariane 5 launcher stands approximately 51.5 m high; the actual figure depends on the upper composite configuration, and the lift-off mass is 740 tons. Lift-off thrust is 11,660 kN. On the ground, the central core vehicle that has an outside diameter of 5.46 m is suspended at the level of the first-stage forward skirt between two solid propellant boosters. This connection, through which the thrust of each booster is introduced in the core stage, is made of alternate elastomer and metallic shims, designed to provide a damping effect on vibrations induced by booster combustion (Fig. 6).

The cryogenic main stage, developed under Aérospatiale as the prime contractor, is 30.5 m, and its dry mass is 12.2 tons. This stage contains 158 tons of liquid hydrogen and liquid oxygen. The Vulcain engine (Fig. 7) is mounted on a thrust cone that distributes thrust evenly at the base of the liquid hydrogen tank. The engine can be oriented along two orthogonal axes by hydraulic actuators operating on the lost fluid principle. This fluid is stored in tanks operating in the blowdown mode. Connecting struts between the main stage and the solid propellant boosters provide rigidity for the rear part. The light alloy propellant tanks have a common bulkhead, insulated with expanded polyurethane. The hydrogen tank has a volume of 390 m³ and is pressurized at values that vary according to the flight phase (between 2.15 and 2.35 bar), using hydrogen tapped at the outlet from the engine regenerative circuit. The oxygen tank (120 m³) is in the upper position and is pressurized with helium gas (3.5 dropping to 2.85 bar), obtained by heating liquid helium in a heat exchanger located in the oxygen turbine exhaust line. This helium (1.15 m³) is stored in a superinsulated tank mounted on the thrust frame. The engine control system is supplied with helium

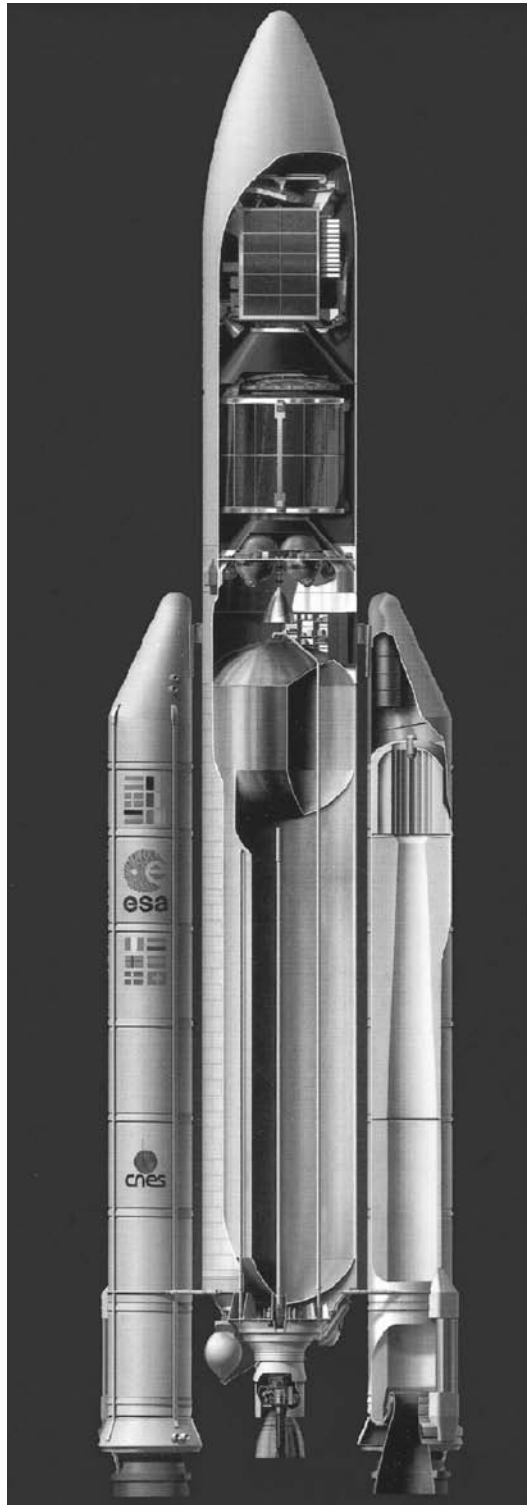
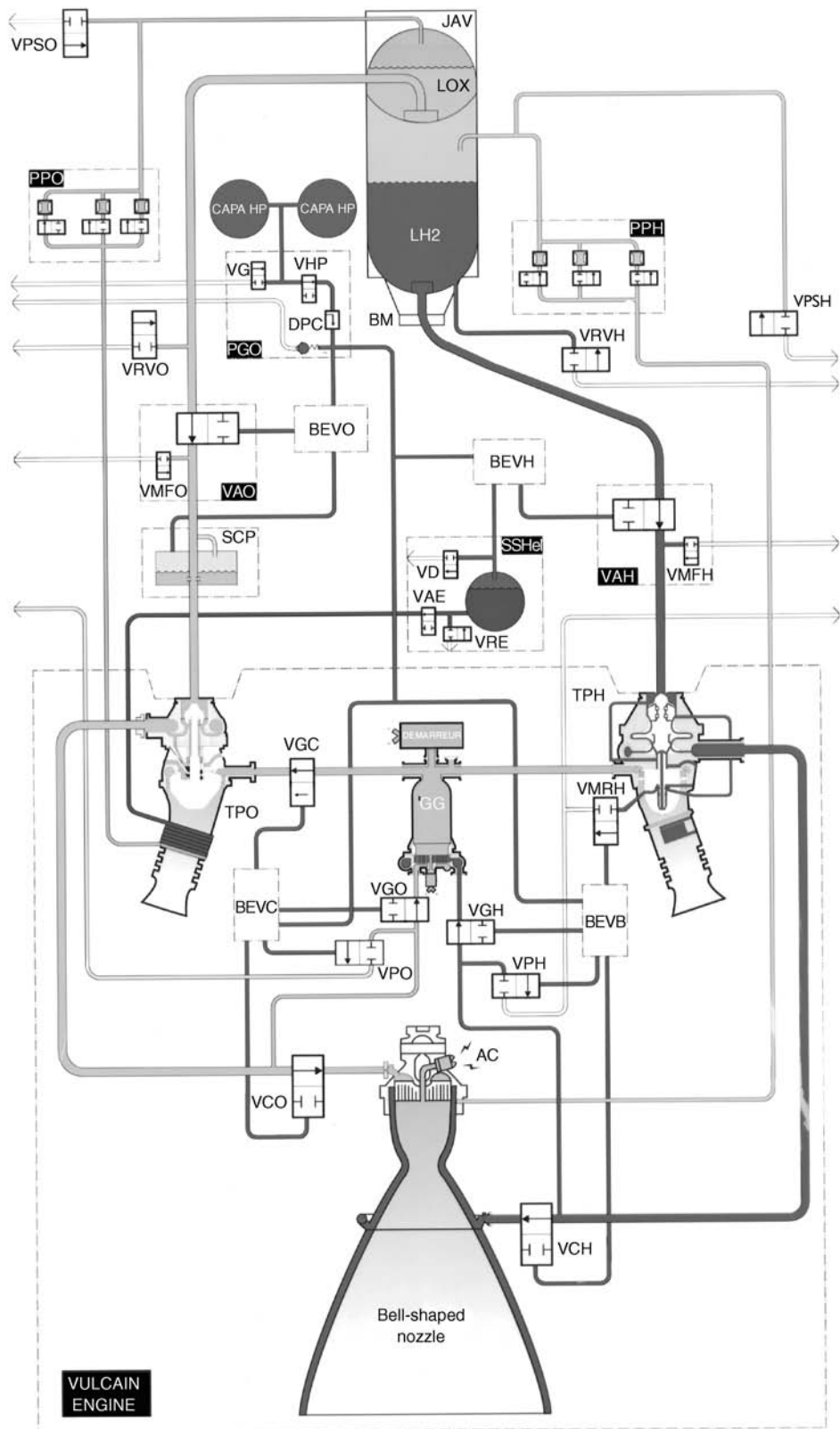


Figure 6. This is a cutaway drawing of Ariane 5. The core is the Vulcain Engine with a payload of two satellites shown above the fuel tanks. The two solid strap-on motors are shown on either side of the core. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



gas stored under pressure in separate tanks. A thrust frame, secured to the upper part of the oxygen tank, receives the thrust of the solid propellant boosters, which is then distributed evenly toward the upper composite.

Europropulsion (joint subsidiary of SNECMA/SEP and FIAT/BPD) was the prime contractor for development of the Ariane 5 solid propellant boosters.

Each booster is 31.16 m high, with a diameter of 3.05 m and a post-combustion mass of 39.3 tons. It contains 238 tons of solid propellant grain, a composite with an ammonium perchlorate and polybutadiene base charged with aluminum. The booster casing is made of a high-strength low-alloy carbon steel and comprises seven cylindrical sections and two bulkheads. The sections are flowturned to a thickness of 8 mm from forged preforms and then assembled using a tang and clevis connection. The sections and bulkheads are assembled to form three segments, each of which is loaded independently with propellant. Internal thermal insulation, made of rubber-based, silica or fiber-filled material, protects the structure from hot combustion gasses. The forward segment is loaded with a 20 tons, star-shaped solid propellant block in Italy. In view of the mass and size of the boosters, a dedicated plant has been constructed in French Guyana for fuelling the central and rear segments (approximately 110 tons of propellant each).

The nozzle, with a flexible bearing made of alternate elastomer and metallic shims, can be steered up to 6° to control the thrust vector. The hydraulic actuators are driven by fluid stored within high pressurized, carbon-fiber vessels operating in blow-down. This fluid is ejected at the nozzle exit. The nozzle, which is highly integrated with the motor, represents a prudent extrapolation of the nozzles developed and qualified for defense applications. The throat in carbon/carbon material ensures minimum erosion during flight. The exit cone is composed of a light alloy housing, with phenolic carbon and silica insulation.



Figure 7. ARIANE 5—Vulcain flow diagram. The Vulcain engine delivers 1140 kN thrust in vacuum and has a specific impulse of 432 sec. It uses conventional gas generator cycle technology. The propellants are delivered by two independent turbopumps. The liquid hydrogen unit operates at 34,000 rpm and comprises a two-stage centrifugal pump, preceded by an inducer that ensures favorable intake characteristics. This pump is driven by a 12-MW, two-stage turbine. The pump delivers up to 560 L/s of liquid hydrogen at a pressure of 17 MPa. The single-stage liquid oxygen turbopump operates at 13,400 rpm and delivers 177 L/s of propellant at 13 MPa (3.7 MW). The two turbines are driven in parallel by a single radial injection gas generator, that operates at a combustion pressure of 8 MPa. The generator is supplied with propellant tapped off at the pump outlet. The combustion chamber pressure is 110 bar. The liquid hydrogen enters the propulsion chamber via an annular distributor. Most of this flow is routed through channels integrated in the double-walled structure of the combustion chamber and throat assembly. The nozzle is cooled by a simple process known as dump cooling: the remaining hydrogen flow is routed through 460 spirally welded iconel tubes, whose diameters increase to give a continuous, bell-shaped surface, then escapes through micronozzles set along the bottom rim of the main nozzle. Although these gases do not undergo combustion, they are heated during the trip and contribute to overall thrust. The turbopumps, gas generator, and combustion chamber ignitions are started by pyrotechnic cartridges. The mixture ratio (mean value 5.25) is adjusted by a two-way valve, used to modify this ratio to terminate combustion on quasi-simultaneous depletion of the two propellants. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The boosters are jettisoned from the core stage by means of pyrotechnic cords. Solid fuel thrusters located in the rear part and in the forward cone are used to distance the boosters from the launcher in a radial plane.

The booster thrust evolves in flight in order to limit general loads, with a maximum of 6,700 kN and an average value of 4,900 kN in vacuum. It must remain within very tight tolerances to limit any thrust differential between the two boosters at any moment. The maximum combustion pressure is 60 bar, and the specific impulse in vacuum is 270 s.

The vehicle equipment bay is 1.56 m high and constitutes a linkage structure between the first stage, second stage, and fairing.

Developed under the prime contractorship of DASA, the second stage is fitted inside the vehicle equipment bay. This is an internal and relatively compact stage, with a diameter of only 3.94 m and a height of 3.36 m. The dry mass is about 1,250 kg. This second stage carries the payload adaptor (Fig. 8).

The propulsion system comprises an “Aestus” engine, burning storable hypergolic propellants (MMH and N2O4), loaded in helium-pressurized tanks and consequently requiring no turbopump. This technological solution was adopted for its reliability, and its simple operation and re-ignition. A total of 6,550 kg of N2O4 and 3,200 kg of MMH (maximum load mass) are contained in two pairs of identical cylindrico-spherical tanks, arranged axisymmetrically two by two. The lift-off mass is 11 tons.

The engine combustion chamber, extended by a nozzle in refractory steel, is cooled by an MMH circuit. The mixture ratio is adjusted to a value of 2.05 by calibration on acceptance testing. The engine can be oriented through an arc of $\pm 4.8^\circ$ on two axes, using electrical servo actuators. The Aestus engine



Figure 8. Ariane 5 second stage. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

develops 29.1 kN thrust and has a specific impulse of 324 s. The burn-time is about 19 min.

Other details of Aestus engine are as follows:

Propellant feed rate: 8.77

Nozzle expansion ratio: 84

Mass: 111 kg.

The bay carries all functional electrical equipment used throughout the mission and a hydrazine attitude control system operating in the "blowdown" mode. This system is used to control roll throughout the propulsive flight, except during the solid booster burn phase, and during the attitude control phase after shutdown of the second-stage engine. By comparison with the Ariane 1 to 4 programs, the functional electrical systems have been fully duplicated and operate in the full active redundancy mode. This might allow for progressive incorporation of standard electrical components.

The vehicle equipment bay is separated from the main stage by dilation of a sealed pyrotechnic tube, which causes a fragile flange to rupture. The upper composite is then distanced by pyrotechnic actuators.

The assembly sequence and facilities in French Guiana (Launch site N° 3 : ELA3) are based totally on experience with Ariane 4, taking due account of the new mass and dimensional values involved, and carrying the policy of stripping the pad down to the bare minimum a step further.

As for the previous Ariane launchers, the synchronized sequence commences at time $T - 6$ min, 30 s. It is subdivided into two separate automatic procedures that involve the fluid systems, on the one hand, and the electrical systems, on the other. At $T - 4.5$ s, an execution command signal is sent to the onboard computer, which then activates the inertial guidance systems and authorizes ignition of the Vulcain engine at time T . Following verification of correct engine combustion at $T + 6.2$ s, the solid propellant boosters are ignited at $T + 7$ s. The boosters burn for about 130 s. They are jettisoned after burnout, when an acceleration limit of about 5.2 ms^{-2} is detected. For a typical GTO launch, booster separation occurs at an altitude of approximately 60 km at about $T + 140$ s. By this time the launcher relative speed has reached approximately 2000 m/s. The fairing is jettisoned on the basis of a thermal flux calculation, at about $T + 3$ min 10 s. Shutdown of the Vulcain engine is initiated by measuring propellant depletion at about $T + 9$ min 46 s. Launcher speed is then approximately 7600 m/s. After separation, fallback of the main stage is controlled so that the stage hits the ocean approximately 1800 km from the Colombian coast. The upper stage is ignited at about $T + 10$ minutes. During the launch, onboard telemetry is transmitted to Kourou via ground stations located at Natal (Brazil), Ascension Island, Libreville (Gabon), Hartbeesthoek (South Africa) and Malindi (Kenya). The upper stage is fully passivated after separation of the payloads to minimize orbital pollution.

The first Ariane 5 flight was a failure due to a software fault. The second flight (502) took place on 3 October 1997. During the second flight, the only anomaly observed was an excessive roll torque at the limit of acceptability, following jettisoning of the solid propellant boosters and during the Vulcain engine burn phase. This fault was corrected very simply by a minor modification to the orientation of the turbine exhaust nozzles.

The third flight demonstrated the exceptional flexibility of the launcher, based on the simultaneous execution of two largely different missions. The first involved placing a reentry capsule into a suborbital trajectory. The precision of the point of impact in the Pacific was excellent. The other mission involved injecting a payload into a geostationary transfer orbit, again achieved with excellent precision and followed by a second-stage maneuver. This mission involved several successive reignitions and shutdowns of the second stage.

Comparison Between Ariane 44 L And Ariane 5

Unquestionably, Ariane 5 marks a technical break with earlier members of the Ariane family. However, continuity in human resources, management principles, and industrial organization ensured that the Ariane 5 program benefited from all experience previously acquired and made it possible to improve the intrinsic reliability by one order of magnitude. The Ariane 5 launcher has four active stages, whereas Ariane 44 L has seven. The number of separation sequences has thus been reduced. Ariane 5 does not have a launch table that is active in positive time, as is the case with Ariane 4. Nor does Ariane 5 have cryogenic fuelling arms that retract at the last moment before ignition and lift-off. Ariane 5 fluid connections are passive and are pulled free on lift-off. The aborted launch case, which requires extremely complex revalidation operations for Ariane 4, is thus simplified to the extreme. Vulcain engine ignition on the ground was deliberately chosen (following ignition failures observed on Ariane 3 flights 15 and 18), as was also the decision to wait to establish a stationary regime for the complete stage (following the flight 36 failure) before checking the engine and authorizing ignition of the solid propellant boosters. The price of this option is a loss of performance (Vulcain engine thrust was set to ensure minimum acceleration of the launcher after boosters were jettisoned. On the other hand, the Vulcain nozzle could have been adapted more efficiently in the case of in-flight ignition). The production of hydraulic power required to operate the booster and first-stage attitude control actuators is obtained from simple pressurized tanks on Ariane 5, compared with the hot-gas-fed engines used for Ariane 4. This list, which covers only simplifications visible at launcher system level, can be extended for each subsystem and component.

Ariane 5 and Hermes

It has frequently been said or written that the Hermes project led to the design of a launcher which was too large, or nonoptimum, in stage configuration. The need to improve performance yet again, to face up to the competition and remain competitive, clearly demonstrates that the first criticism was foundless.

No choice was made to the detriment of optimization for GTO launches. On the other hand, where a number of possibilities were equivalent for this optimization, choices were based on a Hermes criterion. Speaking in 1999, one can be

thankful that this approach was adopted for at least three reasons:

- The Ariane 5 launcher is fully adapted for all types of missions, from constellations to heavy GEO spacecraft and servicing of the Space Station. The 503 flight demonstrated a high degree of flexibility during a single mission.
- The growth potential of Ariane 5 is very substantial, including replacement of the existing upper stage by a cryogenic stage.
- Manned spaceflight represented an excellent stimulus for the construction of launcher reliability. In this area again, the approach adopted was in no way unfavorable for unmanned missions. Abandonment of the Hermes program changed none of the choices made before this decision.

Ariane 5 Evolution Program

The increase in individual mass values of satellites in the *Arianespace* order book continued steadily throughout the 1980s. Extrapolation of this growth demonstrated that Ariane 5 performance was likely to be insufficient by about one ton shortly after qualification of the new launcher. However, this demonstration failed to convince at the beginning of the 1990s. The Hermes spaceplane program was encountering difficulties, including mass-related problems, and the proposal to increase the performance of the Ariane 5 lower composite was consequently regarded as a subterfuge for increasing the size of a launcher, still regarded as too big for launching commercial satellites. Furthermore, the multiplication of constellation projects involving small satellites in LEO and the increased credibility of plasma propulsion techniques for satellite orbit control provided ammunition for those who supported stopping further increases in the mass of geostationary satellites. When the decision to go ahead with the proposed enhancement program was finally made in 1995, this was based more on the desire to preserve cryogenic propulsion expertise than on a genuine need to enhance Ariane 5 performance, even with no change in production costs. The new enhancement program, designated “Ariane 5 Evolution” is aimed at achieving capacity for simultaneous launch of two 3300-kg-class satellites into geostationary transfer orbit.

The following modifications were adopted to meet this specification:

- Increase in Vulcain engine thrust from 1140 to 1350 kN, while changing the mixture ratio from 5.3 to 6, but without modifying the external dimensions of the first-stage tank (although the position of the intermediate bulkhead changed). First-stage mass is increased up to 170 tons of liquid hydrogen and liquid oxygen.
- To recover the specific impulse lost through this modification of the mixture ratio, the turbine exhaust gases are injected into the nozzle, whose expansion ratio area is increased. This technical solution, not adopted for the basic Vulcain program to avoid excessive integration of elements supplied by different contractors, was validated by a technological program in parallel with the main development program.

- Replacement of connections within the three main segments of the solid propellant boosters by welds. In addition to a substantial booster mass gain and lower costs, this simplifies application of internal thermal protection.
- Adaptation of the dual-launch system (SYLDA) used with Ariane 2, 3, and 4 to the dimensions of Ariane 5.

In parallel with the Ariane 5 Evolution program, Arianespace is proceeding with a number of optimization actions (lightening of the vehicle equipment bay, increasing the solid propellant loading of the booster forward segment) based on experience acquired during the development phase. Combined with Ariane 5 Evolution, this will increase the performance in GTO up to 8 tons in single-launch configuration. These enhancements should be operational late mid 2002.

New Phase: Ariane 5 Plus

Changes in the launch service supply situation due to the upcoming availability of Atlas 3, Delta 3, and the American EELVs (Delta 4 and Atlas 5) will further intensify competition and induce a new trend in satellite mass values toward 5 tons for multimedia spacecraft.

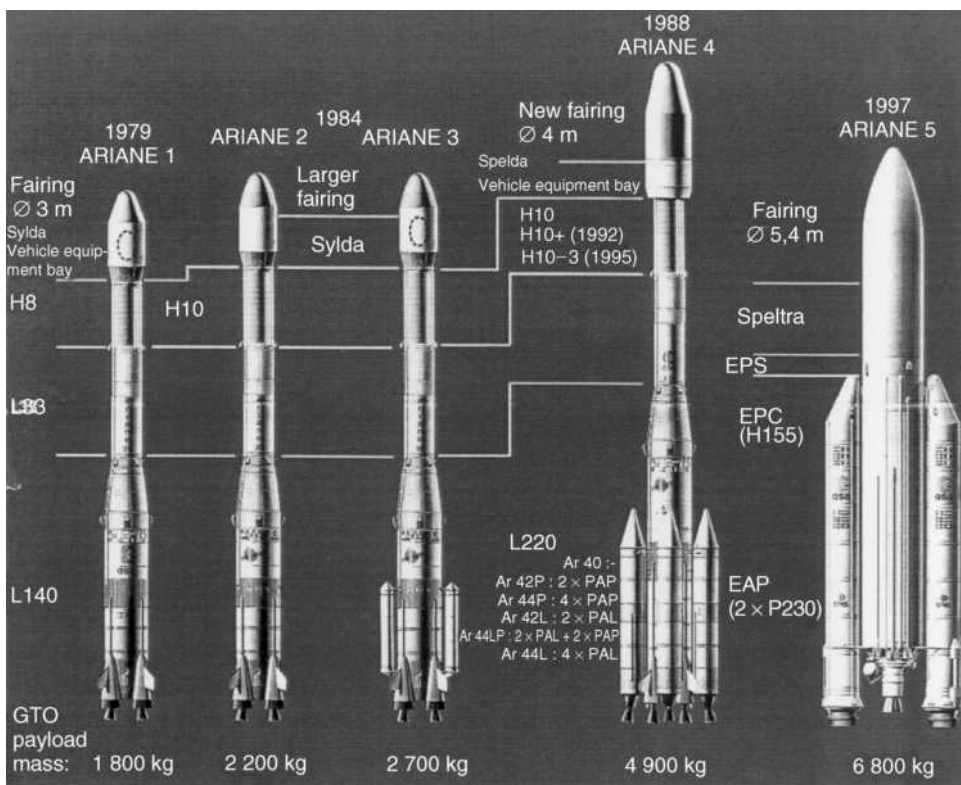


Figure 9. The Ariane rocket family. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

To preserve the competitive advantage achieved with dual launches, the performance of Ariane 5 must exceed 10 tons for GTO injection. This objective will be achieved in two phases:

- For mid 2002 horizon, a second stage will be developed with a propellant load of 14 tons LH₂ and LOX, using the current Ariane 4 third-stage propulsion system. This, combined with Ariane 5 evolution, will take payload capacity up to 10.4 tons in single-launch configuration.
- For a late 2005 horizon, the HM 7 engine will be replaced by the new Vinci motor that delivers 15 tons of thrust and designed for repeated reignition. Tank capacity will be increased to 23 tons of LH₂ and LOX.

This will provide geostationary transfer orbit performance of 11.9 tons in single-launch configuration.

The existing storable liquid propellant upper stage will also be retained but will be adapted to be capable of multiple ignition, with long ballistic phases, to have the flexibility required for all type of missions.

Phase 1 and 2 of this evolution program were decided in the European Space Agency Council in 1998 and 1999. Work on the new engine has been authorized so that the relevant decision can be made in 2001.

Conclusion

Figure 9 briefly summarizes the successive changes in the Ariane launcher and the policy of addressing each new phase takes fullest advantage of experience acquired in the preceding phases. Indeed, Ariane 5 shows a break in this process, and has more the air of the start of a new family. However, the actual break is substantially less pronounced than appears to be the case because the same teams, strengthened by working together for 20 years, have developed and will manufacture Ariane 5.

The technical solutions adopted have frequently been criticized for a lack of initiative because there is sometimes a tendency to confuse efficiency with the technologies applied. But a launcher provides a service, and the only thing that counts is the quality of service provided to the satellite owner. This service is judged in performance, operating cost, reliability, availability, appropriateness of satellite preparation facilities, etc. These are the criteria on which the program's efforts primarily were focused.

BIBLIOGRAPHY

1. *L'Ambition Technologique. Naissance d'Ariane*. P.A.U. Institut d'Histoire de l'Industrie, 1993, pp. 1–91.
2. Fd'Allest. *Interdisciplinary Sci. Rev.* 13 (2): (1988).
3. Ariane 4 User's Manual. Arianespace.
4. Ariane 5 User's Manual. Arianespace.

ROGER VIGNELLES
Corbeil-Essonnes, France

ARTIFICIAL GRAVITY

Definition

Artificial gravity (AG) is not gravity at all. It is not a field force or a “force at a distance.” Neither does its strength obey the inverse square law of attraction that determines the orbital motion of planets. However, in terms of its action on any mass, it is indistinguishable from “real gravity.” Instead of gravitational pull, it exerts a centrifugal force, proportional to the mass that is being accelerated centripetally in a rotating device. Although the effect of AG on an extended body differs from that of true gravity, the effects on any given mass are equivalent. Thus AG is simply the imposition of acceleration on a body to recover the forces that are eliminated by the free fall of orbital flight. (Of course, real gravity is not eliminated in orbit. The pull toward Earth in Earth orbit and toward the Sun in interplanetary orbit is balanced by the “free fall” acceleration of the spacecraft and its contents toward Earth or the Sun. To an observer or instrument onboard the spacecraft, it *feels* as though the pull of gravity were removed.)

Provision of AG

In principle, AG could be provided by various means. A continuously thrusting rocket that accelerated a spacecraft halfway to Mars would generate AG equal to the acceleration level. Intermittent impulsive AG would be imposed on an astronaut who jumps back and forth between two opposing trampolines or even between two stationary walls in a spacecraft. However, the term *artificial gravity* is generally reserved for a rotating spacecraft or a centrifuge within the spacecraft. Every stationary object within the centrifuge is forced away from the axis of rotation toward the outer “floor” by a force proportional to the mass of the object, its distance from the center of rotation, and the square of the angular velocity of the device.

Why AG May Be Necessary

Probably the most serious health threat to humans during interplanetary flight comes from radiation exposure en route and on some extraterrestrial surface. Beyond that, prolonged exposure to weightlessness itself can result in deconditioning many of the body’s systems. For space voyages of several years, such as those envisioned for exploration of Mars, the human requires some sort of “countermeasure” to reduce or eliminate this deconditioning. Intensive and sustained exercise on a treadmill, bicycle, or rowing machine was used on the U.S. and Russian spacecraft to minimize the problems of weightlessness. The procedure is uncomfortable and excessively time-consuming for most astronauts. Furthermore, its effectiveness is not proven for all users. Other kinds of countermeasures, including diet, fluid loading before reentry, lower body negative pressure, or wearing a “penguin suit” to force joint extension against a resistive force are either marginally effective or present an inconvenience or hazard.

The *physiological effects* of weightlessness are generally adaptive to space flight and present a hazard only upon return to Earth or landing on another planet (1). However, they may present hazards in flight in the event of a bone fracture, a vigorous muscle contraction, or alterations in the heart's rhythm.

Aside from the severe danger of space radiation, the principal physiological risk of long flight is deterioration of the skeleton. Bones are living tissue, constantly being strengthened by calcium extracted from the blood and destroyed by returning calcium to the blood. Bone maintenance requires a compressive load along the axis of the bone and some high-force impulsive loading. In the absence of these loads that are normally provided by gravity and walking, the major bones that support body weight begin to deteriorate, and a net loss of body calcium occurs, independent of the amount taken in with food or supplements. The long bones in the legs and the vertebrae in the spine lose crucial size and strength during prolonged bed rest. Similarly, they lose strength in spaceflight. Calcium is lost at a rate of about 1/2% per month, and the losses are reflected in the density and size of weight-bearing bones. For a spaceflight of two years, a 25% decrease in bone size might occur (unless the process reaches a plateau), thus increasing the risk of fracture and severely hampering the bone's ability to mend.

Muscles involved in weight bearing, as well as bones, begin to weaken with disuse in weightlessness. The major muscle groups in the legs and back that normally support weight lose mass and are also "reprogrammed," so that fibers previously devoted to slow steady tension are used for brief bursts instead. The shifting of fluid from the legs and lower trunk to the head and chest that produces the first symptoms of head-fullness discomfort on orbit initiates an early loss of body fluid, including blood plasma. The relative excess of red blood cells is countered by stopping their production in the bone marrow and additionally by destroying young red blood cells. The cardiovascular regulating system that acts to maintain adequate blood pressure when we stand up, is no longer needed in space and shows signs of deterioration. Neither the fluid loss and resulting "space anemia," nor the loss of cardiovascular regulation and tone normally cause any difficulty in orbit. During reentry and back on Earth, however, the renewed exposure to gravity can cause weakness and fainting.

The balance system that keeps humans from falling depends on the detection of gravity by the otolith organs in the inner ear. Because the only stimulus to the organs in weightlessness is linear acceleration, considerable reinterpretation of vestibular signals takes place. A consequence of this process is the common occurrence of space sickness early in flight and postural disturbances and vertigo after return.

The immune system that fights infection may also be compromised by space flight, although it is unclear whether weightlessness alone is the major factor.

In addition, a variety of *human factor* problems arise in weightlessness, including the constant need for handholds or footholds for stabilization and the possibility of disorientation within a spacecraft. However, these problems are often balanced by the ease of moving heavy objects, the use of three-dimensional space, and the shear pleasure of floating in weightlessness.

History of AG

The notion of creating a substitute for gravity through centrifugation was introduced early in the conception of human space travel. Tsiolkovsky, the influential Russian space visionary, discussed the idea in 1911, and his concepts were picked up 50 years later by Korolev, who designed a flexible tether system for the Voskhod manned missions (2). It was, however, never built. A detailed engineering proposal for an AG station was introduced by Noodhung in 1927, a full 50 years before the first satellite was launched. When Von Braun described his vision of space exploration in 1953, he included a large rotating torus to deal with weightlessness (Fig. 1) (3).

The popularization of AG, however, is attributable to the science fiction community. The large rotating torus in Clarke and Kubrick's *2001: A Space Odyssey* presented an idealized version of life in space, free of health problems and the negative effects usually associated with transiting from the rotating to the stationary parts of the station. By 1965, preliminary tests on a short-radius centrifuge first showed that subjects who were deconditioned by bed rest could be protected against cardiovascular deconditioning by periodic centrifugation (4).

Experience with AG in space has been quite limited. Rats were centrifuged continuously at 1 g for several days and showed no deconditioning. Human experiments, however, have not been conducted to date. Early attempts to test AG by tethering a Gemini spacecraft to an Agena rocket were inconclusive and nearly led to disaster when the thruster nozzle stuck on Gemini 8, sending the pair of space vehicles into an uncontrollable spin. The 2.5-m-radius centrifuge on the International Space Station should afford the opportunity to examine the adequacy of various levels of AG in protecting rodents during spaceflight.

Design Boundaries

The envelope of operation for AG is limited by several factors, as pointed out by Von Braun and adapted by others. The “comfort zone” for AG with a rotational radius of up to 1000 feet is bounded by several constraints (5). In one presentation, the nominal design point was for a 734-foot radius architecture, spinning

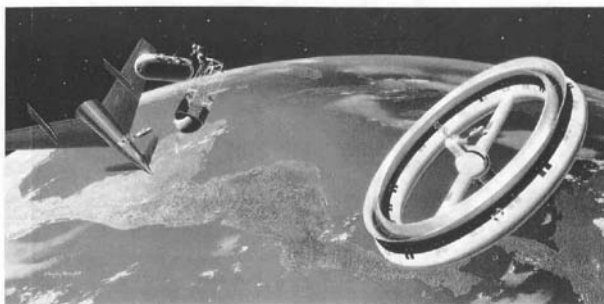


Figure 1. Von Braun's rotating space station (reprinted from Reference 3). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

at approximately 1.8 rpm. This large radius creates less than a 1% gravitational gradient from head to foot under a 4% ratio of Coriolis force to apparent weight for a crew moving at 3 ft/s. The rim velocity would exceed 200 ft/s. The basic design space is normally shown on a graph of rotational rate versus radius; the acceleration level appears as a derived parameter according to the equation $A = r\omega^2$. The design boundaries have generally been stated for continuous rotation but are also shown here for intermittent centrifugation (Figs. 2a,b).

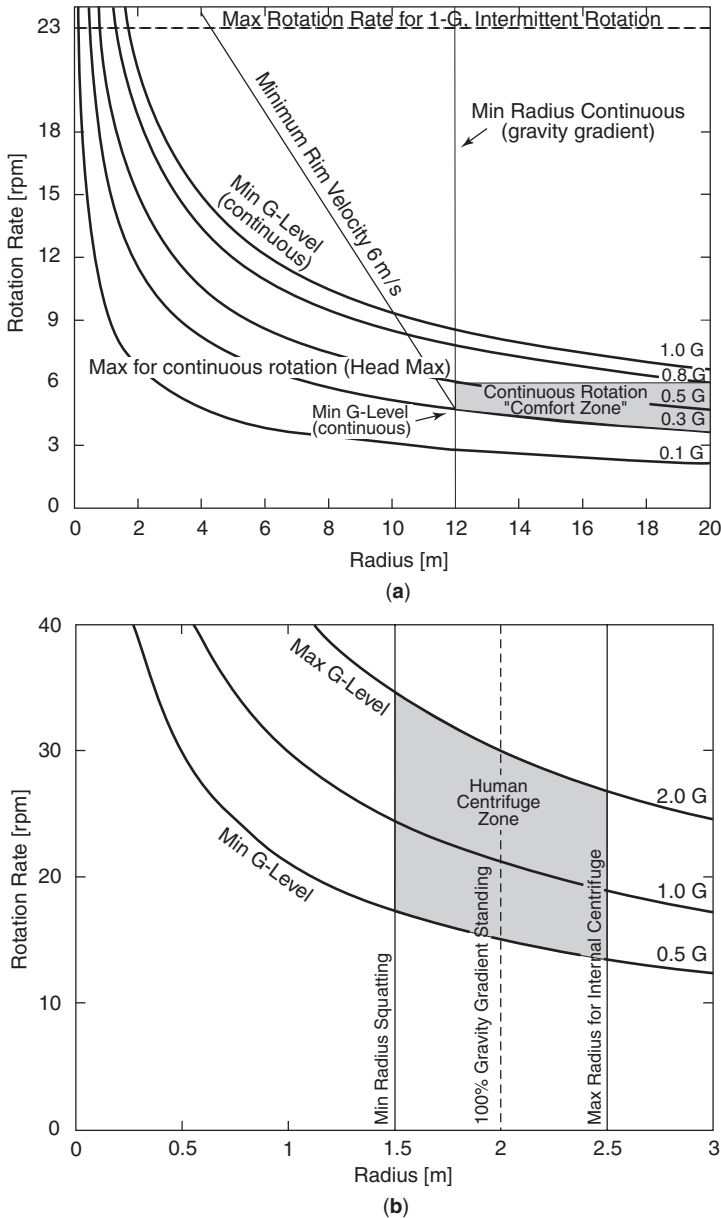


Figure 2. (a) Comfort zone for continuous AG (adapted from Reference 5). (b) Centrifuge zone for intermittent AG (adapted from Reference 5).

The *minimum gravitational level*, normally measured at the rim of a centrifuge, is the key parameter in the design space. The limited animal tests in orbit confirm that continuous rotation to yield 1 g at the feet of a small rodent is sufficient to maintain normal growth. However, it remains to be determined whether a lesser g level will suffice. Based on centrifuge studies of long duration, Russian scientists suggest that the minimum level of effective AG is about 0.3 g and recommend a level of 0.5 g to increase a feeling of well-being and normal performance.

The *maximum gravitational acceleration level* is also a factor if short-radius intermittent AG is used. Levels up to 2 g are probably useful, especially if combined with exercise, but a level as high as 3 g's is likely to produce ill effects if maintained for more than 20 minutes.

The *maximum angular velocity* of the AG device is limited by the Coriolis forces encountered when walking or when moving objects, and by the motion sickness and disorientation experienced with certain kinds of head movements. Coriolis accelerations are real inertial accelerations that occur when moving within a rotating framework. Any movement in a straight line with respect to the rotating frame, except for one parallel to the axis of rotation, is in fact a curved motion in inertial space. The curve reflects acceleration sideways and entails a sideways inertial reaction force (Fig. 3).

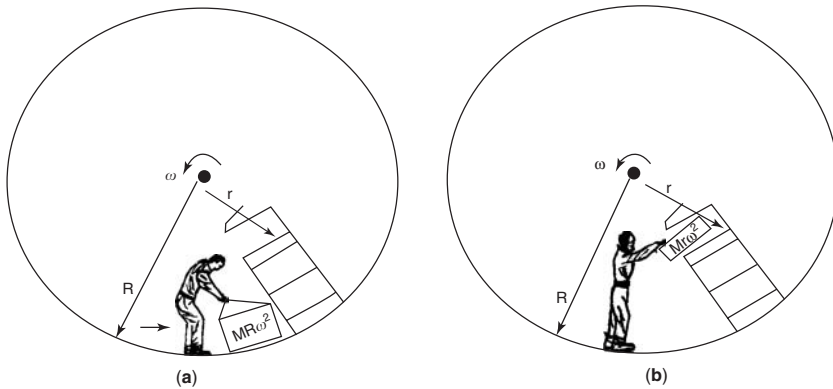
People trying to walk radially outward on a spinning carousel will feel a surprising force pushing them sideways, parallel to the circumference. As seen by an observer stationed outside the carousel, the walker's path is really curved in the direction of the carousel's spin. The sideward inertial acceleration requires a sideward force (*Coriolis force*), according to Newton's second law, and the subjects need to apply that unexpected force to avoid walking a path that is curved relative to the carousel. They also must apply an unexpected postural reaction to avoid falling over.

Additionally, anyone trying to walk along the rim of the AG spinning vehicle in the direction of the spin is subject to an unexpected radial inertial acceleration inward, which entails a downward Coriolis force, making the space walker feel heavier. If the astronaut were to turn around and walk along the rim in the direction opposite to the spin, the Coriolis force would be upward and the apparent weight of the astronaut would be reduced. The magnitude of the Coriolis force is given by the equation $|F_c| = |m\omega \times v|$. From considerations of human factors, the Coriolis accelerations should be kept to less than some fraction of the AG gravity level. Stone (5,6) suggests that this be no higher than one-fourth. For radial movement at velocity v , this is given by

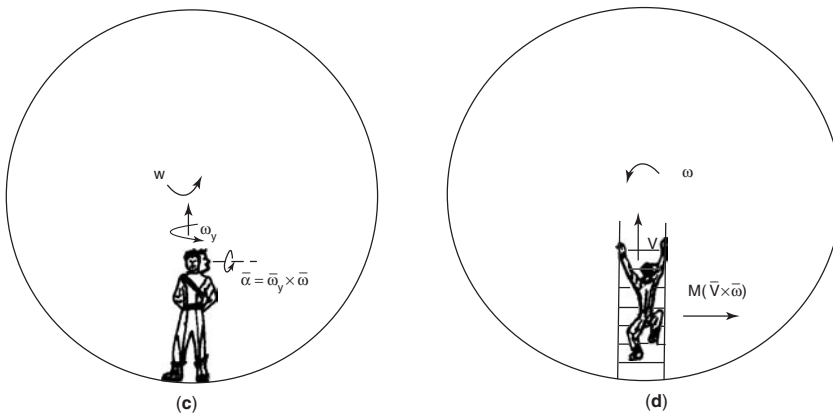
$$\text{Coriolis acceleration/artificial gravity} = \omega v_{\max}/r\omega^2 = v_{\max}/v_{\text{rim}},$$

where v_{rim} is ω times r and is the speed of the outer rim of the AG centrifuge. The *minimum rim velocity* is limited only by the need to maintain enough friction for locomotion when walking against the direction of spin. For walking, v_{\max} is about 1 m/s, and it has been assumed that the estimated minimum rim velocity is 6 m/s.

The most disturbing aspect of AG rotation is probably the Coriolis cross-coupling accelerations detected by the semicircular canals in the vestibular systems of the inner ear. The organs function to detect angular velocity of the

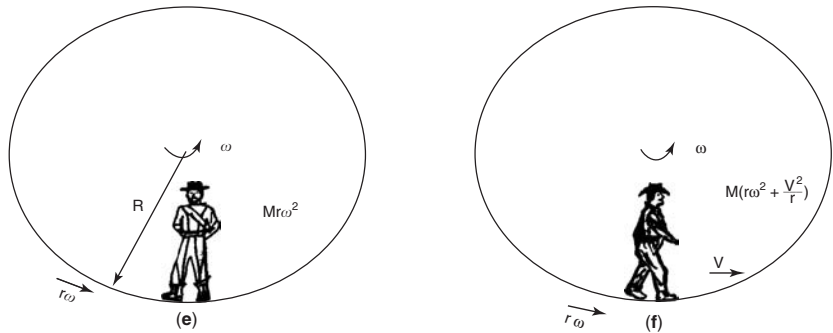


Gravity gradient. (a) Object of mass M "weighs" $MR\omega^2$ on the rim. (b) Object of mass M "weighs" $Mr\omega^2$ at radius r .



*Cross-coupled acceleration. Yaw head velocity produces acceleration about the naso-occipital axis.

*Tangential Coriolis forces. Forces push sideways on an astronaut moving radially in the spinning spacecraft.

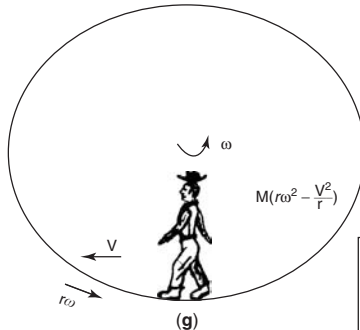


*Radial Coriolis forces during locomotion.

(e) Standing.

(f) Moving in direction of spin.

(g) Moving in direction opposite of spin.



*Adapted from Fig. 8, "Influences of artificial gravity on locomotion." (caption), in R.W. Stone, "An overview of artificial gravity," 5th Symposium on the Role of Vestibular Organs in Space Exploration, Pensacola, FL, Aug. 1970 (NASA SP = 314 : 28).

Figure 3. Coriolis forces (adapted from Reference 5).

head relative to inertial space for most normal head movements. However, because of their mechanical structure, they fail to register long-lasting constant velocity motion and, instead, indicate that one is stationary in a turn that lasts more than 10–20 s. In AG, these vestibular signals are apparently inconsistent with what one sees in the spacecraft and also with the linear acceleration registered by the otolithic organs in the labyrinth. This conflict, before adaptation, produces both motion sickness and spatial disorientation.

When subjects in AG suddenly move their heads about an axis that is not parallel to the spin axis, two unexpected angular accelerations occur. First, during the head movement a “cross-coupled acceleration” occurs, equal to the product of the spin rate and the head angular velocity that produces transient acceleration about a third orthogonal axis. This is given by the equation $A = \omega_y \times \omega_s$. Second, when the head is turned, the spin angular velocity is moved from one head plane to another, producing a sensation of deceleration about the first axis and acceleration about the second one. A sensation of rotation with components around both axes usually occurs for up to 10 sec, as the semicircular canals return to their neutral position. The directions of both the Coriolis force and the cross-coupled accelerations depends on the direction the subject is facing in the rotating spacecraft, as well as the direction of head movement, thereby complicating the process of general adaptation to the unusual environment.

All of the unexpected Coriolis sensations are proportional to the AG spin rate. Although further adaptative schedules might increase the tolerable rate, the *maximum spin rate* for continuous rotation has been estimated at 6 rpm, with possible elevation to 10 rpm. Almost all subjects can adapt quickly to work in a 2-rpm rotating environment. It is believed that most could tolerate increased rotational rates to 6–10 rpm, providing that they are built up slowly in steps of 1–2 rpm with a period of 12–24 h at each increment (7).

The *gravitational gradient* refers to the change in AG level with radius and can affect both physiological function and the ease of handling materials in space. Since the “g level” is proportional to the radius, the gravitational gradient from head to foot is simply the ratio of height to radius: $Gradient = h/R$. For continuous rotation at smaller radii, comparable to the astronaut’s height, the gravitational gradient may become more of a problem. For a 2-m astronaut, the radius would be at least 4 m for a 50% maximum gradient.

Experimental Results

Space Experiments. Despite the long-standing interest in artificial gravity, experimental evidence from space is very limited. Only two space missions early in the space program were devoted to animal studies, and all of the human in-flight results were anecdotal.

Flight Animal Experiments. The Soviet space research community expressed an early and intense interest in AG and, in 1961, began testing rats and mice in the 25-s weightless periods of parabolic flight. Animals showed normal appearing locomotion during these brief periods if they were housed in a centrifuge producing at least 0.3 g, thus setting this as a minimum g requirement (8). The first animals to be centrifuged in space were on the Cosmos 782 mission

in 1975, when fish and turtles centrifuged at 1 g were found indistinguishable from their ground controls. Furthermore, turtles centrifuged at levels as low as 0.3 g showed none of the muscle wasting typical of weightlessness. A much more extensive investigation was carried out on rats centrifuged during the 20-day mission of Cosmos 936 in 1977. These animals, housed in a small-radius (32-cm), high-speed (53.5-rpm) 1-g centrifuge, showed deficits in equilibrium and postural control postflight, consistent with the observed reduction in vestibular sensitivity. Faring less well than their ground controls, they also failed to counter fully the usual effects of weightlessness on loss of muscle and bone, circumstances that may have been the result of the small cage size and the high-g gradient. The large animal centrifuge planned for the International Space Station is designed to provide a range of AG levels, above and below 1 g, to a large variety of fish, plants, and small animals.

Human Space Experience with AG. No formal human AG experiments were performed in space during the first 40 years of the space age. During the earliest years of human spaceflight, the major physiological disturbances involved “space adaptation syndrome” and were of concern only for the first few days in orbit. The debilitating effects of weightlessness on the bone, muscle, and cardiovascular system were demonstrated on the Skylab missions in the early 1970s and later on the long-duration Salyut and Mir flights. However, it was believed that in-flight exercise, augmented by resistance training and fluid loading, would solve the problem. As time passed, the opportunities for human centrifuges or rotating spacecraft in orbit disappeared. During a 1966 Gemini mission, an orbiting Agena rocket casing was tethered to the spacecraft, and the two were put into a slow spin. No data were taken. On Gemini 8, when Gemini was docked to the Agena, a planned slow rotation got out of control because of a stuck thruster, and the crew was saved only by the skillful use of an orbital maneuvering engine. No further spacecraft AG tests have been conducted. Since then, the only opportunities for investigation have come from uncontrolled, anecdotal reports.

During the Skylab missions, the crew took advantage of the large open compartment to run around the curved circumference. They produced a self-generated AG by running. The crew reported no difficulty with either locomotion or motion sickness.

Although no specific AG human experiments have been performed, some centrifugation for other purposes has produced a measure of centripetal acceleration. During the Spacelab International Microgravity Laboratory (IML-1) mission, subjects were spun on a rotator in which the head was 0.5 m off center, experiencing an acceleration of $-0.22 g_z$, and the feet were on the other side of the axis, experiencing an acceleration of $+0.36 g_z$ (9). No unusual inversion phenomena were reported. Similarly, in the Neurolab Spacelab mission, four subjects received periodic rotation in a similar situation without reorientation. In that case, however, those subjects seemed to have achieved some measure of resistance to postflight orthostatic instability and did not show the usual decrease in vestibular sensitivity to tilt (10).

Ground Centrifuge Experiments. Despite the absence of flight-test opportunities, several laboratories worldwide have continued ground-based studies of the efficacy and acceptability of human horizontal centrifugation. Of course, all of these investigations are hampered by the presence of the steady gravitational

pull. Gravity adds to the centrifugal force vectorially and produces a net specific gravito-inertial force $F = g - a$ directed between vertical and horizontal.

The earliest of the extensive tests of sustained rotation were conducted in Pensacola (11), beginning in 1958. The "slow rotating room" (SRR) exposed volunteers to prolonged rotation (12). This 3-m-radius room that has a horizontal floor permitted subjects to adapt to rotation during several days (13,14). Initially, most subjects developed motion sickness symptoms when they made head movements at room rotational rates in excess of 3 rpm and, through that experience, learned to restrict them. Incremental increase in the speed of the room was employed. After several days, most subjects were able to make head movements without symptoms at rotational rates up to 6 rpm. Only some of the subjects could go further to move comfortably at 10 rpm. When the rotation was stopped, subjects felt an aftereffect and an erroneous motion sensation during head movements. They were maladapted to rotation in the opposite direction.

Beginning in the 1960s a major ground research program on AG was conducted at the Institute for Biomedical Problems in Moscow (IBMP). Their earliest tests in the MVK-1 small rotating chamber at speeds up to 6.6 rpm allowed rotating one or two subjects for up to a week. It was followed by the roomier 10-m-radius "Orbita" centrifuge, capable of rotating two to three people for several weeks at speeds up to 12 rpm. The longest tests were for 25 days at 6 rpm. The initial exposures produced the expected disturbance of equilibrium and coordination. Within an hour, the usual pattern of motion sickness symptoms occurred, including vomiting in some cases (15). In 4–5 hours, subjects also complained of listlessness, sleepiness, and headache—similar to the Sopite syndrome identified by Graybiel. Three periods of vestibular adaptation were distinguished for these long-duration exposures. The first 1–2 days were characterized by severe motion sickness. This was followed by a week during which the nausea and related acute symptoms disappeared, but listlessness and headache remained. Finally, after the first 7–10 days, subjects showed immunity to motion sickness, even when additional vestibular stimulation was imposed. The generalizability of this adaptation has not been determined. The Soviet centrifuge tests indicated an absence of any motion sickness symptoms at 1 rpm, moderate symptoms at 1.8 rpm, and marked symptoms at 3.5 rpm. Head movements brought on discomfort in all cases.

More recent investigations have assessed the ability of subjects to avoid motion sickness during head movements while rotating at the high speeds associated with short-radius centrifugation. Antonutto and colleagues in Udine, Italy, found that subjects who were pedaling on a bicycle-powered short centrifuge were able to make head movements without acute motion sickness while rotating at 19–21 rpm. Young, Hecht, and colleagues used the 2-m-radius centrifuge at MIT to show that most subjects could adapt both their eye movements and their motion sickness symptoms to accommodate head movements while rotating at 23 rpm (16). Both the Udine and the MIT studies were conducted at speeds sufficient to produce 1 g of horizontal centripetal acceleration or a net gravito-inertial acceleration of 1.4 g's. In the Udine centrifuge, it was aligned with the subject's, head-to-foot axis, whereas in the more provocative MIT studies, the subject remained horizontal.

The Coriolis forces associated with limb movements, head movements, and walking in a rotating environment are initially both surprising and disturbing.

However, in almost all cases, appropriate new motor control strategies are developed, so that subjects can adapt to the new environment and no longer are even aware of the unusual forces. Extensive experiments in the Brandeis University rotating room demonstrate the remarkable ability to adapt to unusual environments (17). A measure of dual adaptation apparently exists, so that subjects can switch from the rotating to the nonrotating environment with minimal relearning.

The adequacy of artificial gravity in stimulating the cardiovascular system has been investigated in ground studies. In most studies, the debilitating effects of weightlessness are simulated by sustained bed rest, often at 6° of head-down tilt and occasionally by partial submersion in water to approximate the fluid shift better that occurs in space. In a pioneering study in 1966, White and his colleagues at Douglas (18) showed that intermittent exposure to 1g or 4g's on a 1.8-m-radius centrifuge was effective in alleviating the usual decrease in tolerance to standing (orthostatic intolerance). Exercise produced little additional benefit. The principal cardiovascular reactions of interest for centrifugation are the venous tone, especially in the legs, and the baroreflex regulation of blood pressure. For a short-radius centrifuge small enough to accommodate a subject only in a squatting position, the centrifugation does little to encourage venous return by stimulating the muscles. The IBMP ground centrifuge tests (19) demonstrated that subjects who were deconditioned by 2 weeks of water immersion could increase their post-immersion tolerance to +3 g_z by intermittent acceleration on a 7-m-radius centrifuge. For some time, it was debated whether the intermittent centrifugation conditioned only the passive motor tone or whether the body's active baroreflex to counter the effects of gravity on blood pressure was also affected. Burton and Meeker (20), using a 1.5-m-radius centrifuge intermittently, showed that the baroreceptors are adequately stimulated during AG. Their slow compensation for the hydrostatic pressure drop during rotation permits the g tolerance to gradual onset acceleration to exceed that to rapid onset acceleration. Beyond even the benefit of intermittent acceleration on cardiovascular responses is the effect on blood volume. Normally, weightlessness or head-down bed rest produces a fluid shift toward the head that in turn leads to fluid loss, including plasma, and a resulting increase in hematocrit. However, Yajima and his colleagues from Nihon University School of Medicine in Tokyo (21) showed that 1 hour per day of 2g_z exposure of their subjects, using a 1.8-m-radius centrifuge, was sufficient to prevent hematocrit from increasing during a 4-day bed rest period. In other studies, they confirmed the effectiveness of intermittent centrifugation on maintaining baroreflex and parasympathetic activity (22). To prevent motion sickness, the Nihon investigators stabilized the head during these centrifuge runs.

The interaction between the cardiovascular fitness enhancement of regular exercise and the tolerance built up during centrifugation remains unclear. Certainly the two countermeasures are individually effective, but whether they contribute more in combination is still under study (23,24).

Artificial Gravity Design Options

The choice of AG design depends on a basic decision whether the crew is to be transported with *continuous* AG, requiring a large-radius device, or exposed to

intermittent AG, in which case a small rotator can be employed. The classical large *spinning space station*, as epitomized by the von Braun torus, was the basis for early designs in the Apollo era (25). At one time, a large toroid 150 feet in diameter and constructed of six rigid modules joined by an inflatable material, was envisioned. The large mass and excess volume of a torus or hexagon forced consideration of other ways of generating centrifugal forces at large radii. The two that emerged are the rigid truss, or boom, and the tether concept. A *rigid truss* design typically would have the crew quarters and operations module at one end and a large counterweight at the other end. The counterweight might be an expended fuel tank or an active element such as a nuclear power source. In most cases a counterrotating hub is present at the center of rotation to provide both a nonspinning docking port and to allow for a zero-g workspace for experiments. A variation on the rigid truss is the extendable or telescoped boom concept, in which the radius of the AG systems could be varied more easily than with a fixed truss and slider. However, both of these designs imply considerably more mass and power requirements than a tether system. A variable length *tether* that could be unreeled in orbit and used to connect a spacecraft to a counterweight has emerged as the most acceptable design for a large AG system. As envisioned for a Mars mission (26), it would consist of a 80,000 kg habitat module 225 m from the center of mass, with a 44,000 kg counterweight 400 m beyond. The two are connected by a tether, weighing 2400 kg, reeled out by a deployer weighing 1700 kg. All told, the additional weight for accommodating a tethered AG system for a human Mars mission is about 21,000 kg, or about 5% of the 0-g weight, plus about 1400 kg of propellant (Fig. 4).

One of the obvious concerns about a tethered AG system is its vulnerability to tether breakage. For the Mars mission design, a tether in the form of a band $0.5\text{ cm} \times 46\text{ cm} \times 750\text{ m}$ would provide a dynamic load safety factor of 7, offering a working strength of (630,000 N). That concern has otherwise been addressed by using webbing or braided cable to maintain tether integrity, even in the event of a meteoroid collision. (The probability of tether impact with a micrometeoroid of mass greater than 0.1 gm was calculated as .001 for a mission of 420 days.) A second concern about a tethered system lies in its dynamic stability, especially during unreeling and during spin up and spin down. The interaction with orbital maneuvers is complex, whether the spin axis is inertially fixed or tracking the Sun to facilitate the use of solar panels.

The alternative approach to AG is to use a *short-arm centrifuge* intermittently. In this case, the exposure would not be limited to less than 1 g, but might be as high as 2 or 3 g's to deliver adequate acceleration in exposures of perhaps 1 h daily or several times per week. Of course, such a short device would have to spin much faster than the 6 rpm limit envisioned for a large continuous system—and would produce significant Coriolis forces and motion sickness stimuli if the head is moved, at least until adaptation occurs. The short-radius centrifuge becomes particularly attractive when its dimensions shrink to the point that intermittent centrifugation could be carried out within the confines of a spacecraft, rather than entailing rotation of the entire complex. A 2-m-radius AG device permits subjects to stand upright and even walk within its limited confines. Of course, the head is then close to the center of rotation, and a significant gravitational gradient appears as one goes from head to toe. Many of the ground

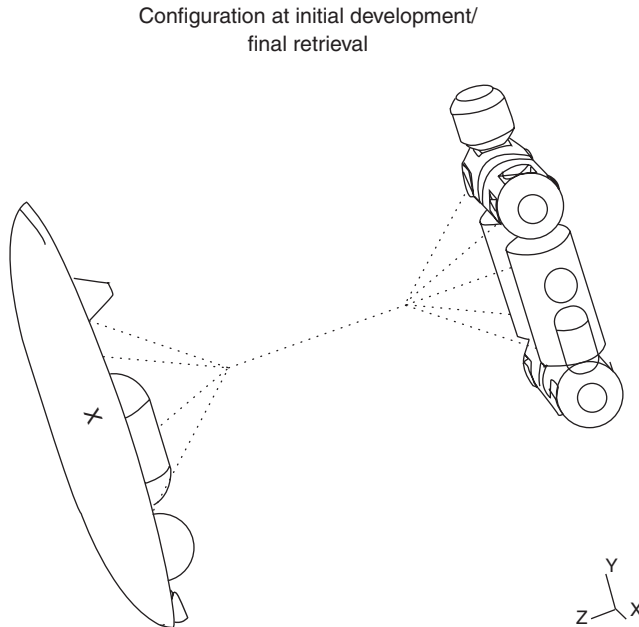


Figure 4. Tether design. (Reference 26).

studies of intermittent short-radius centrifugation have been conducted with rotators of radius from 1.8–2.0 m. As the radius shrinks even further to less than 1.5 m, the taller subjects can no longer stand erect but must assume a squatting or crouching posture. For many such designs, the subject would also provide the power to turn the device and perform valuable exercise by bicycling the centrifuge into rotation (Fig. 5).



Figure 5. Bicycle-driven centrifuge courtesy of Dava Newman (photo by William Litant, Massachusetts Institute of Technology). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Although the power saving may be trivial, or not even used, the importance of active exercise while exposed to intermittent centrifugation might lie in its protection against syncope, or fainting, as the body is exposed to the unaccustomed footward forces that tend to pool blood in the lower extremities.

ACKNOWLEDGMENT

Supported in part by the National Space Biomedical Research Institute through a Cooperative Agreement with the National Aeronautics and Space Administration (NCC 9-58).

BIBLIOGRAPHY

1. Young, L.R. Artificial gravity considerations for a Mars exploration mission. In B.J.M. Hess & B. Cohen (eds), *Otolith Function in Spatial Orientation and Movement*, 871 New York Academy of Sciences, New York, pp. 367–378, 1999.
2. Harford, J. *Korolev*. Wiley, New York, 1973, pp. 7, 13, 186.
3. Von Braun, W., F.L. Whipple, and W. Ley. *Conquest of the Moon*, C. Bonestell and C. Ryan (eds), Viking, New York, p. 11, 1953.
4. White, W.J., J.W. Nyberg, P.D. White, R.H. Grimes, and L.M. Finney. *Biomedical potential of a centrifuge in an orbiting laboratory*. Douglas Report SM-48703 and SSD-TDR-64-209-Supplement. Douglas Aircraft Co., Inc., Santa Monica, CA, 1965.
5. Stone, R.Q.W., Jr. An overview of artificial gravity. In A. Graybiel (ed.), *5th Symposium on the Role of the Vestibular Organs in Space Exploration*, SP-314, 23-33. NASA, Washington, DC, 1970.
6. Faget, M.A., and E.H. Olling. Orbital space stations with artificial gravity. In A. Graybiel (ed.), *Fifth Symposium on the Role of the Vestibular Organs in Space Exploration*, SP-314, 23-33. NASA, Washington, DC, 1970.
7. National Aeronautics and Space Administration, Office of Manned Space Flight. Artificial Gravity Experiment Definition Study. Advanced Manned Missions, 1970.
8. Yuganov, Ye. M. Physiological reactions in weightlessness. In V.V. Parin (ed.), *Aviation Space Med.* NASA TT F-228, NASA, Washington, DC, 1964.
9. Benson, A.J., F.E. Guedry, D.E. Parker, and M.F. Reschke. Microgravity vestibular investigations: Perception of self-orientation and self-motion. *J. Vestibular Res.* 7: 453–457 (1997).
10. Moore, S.T., G. Clement, T. Raphan, I. Curthoys, I. Koizuka, and B. Cohen. The human response to artificial gravity in a weightless environment: Results from Neurolab centrifugation experiments. In M.S. El-Genk (ed.), *Space Technology and Applications International Forum–2000*. American Institute of Physics, 2000.
11. Guedry, F.R., R.S. Kennedy, C.S. Harris, and A. Graybiel. Human performance during two weeks in a room rotating at three rpm. *Aerosp. Med.* 35: 1071–1082 (1964).
12. Kennedy, R.S., and A. Graybiel. Symptomatology during prolonged exposure in a constantly rotating environment at a velocity of one revolution per minute. *Aerosp. Med.* 33: 817–825 (1962).
13. Reason, J.T., and A. Graybiel. Progressive adaptation to Coriolis accelerations associated with 1 rpm increments in the velocity of the slow rotation room. *Aerosp. Med.* 41: 73–79 (1970).
14. Clement, G., S.T. Moore, et al. Perception of tilt (somatogravic illusion) in response to sustained linear acceleration during space flight. *Exp. Brain Res.* 138: 410–418 (2000).

15. Kotovskaya, A.R., R.R. Galle, and A.A. Shipov. Soviet research on artificial gravity. *Kosm. Biol. Aviakosm. Med.* 2: 72–79 (1981).
16. Young, L.R., H. Hecht, L. Lyne, K. Sienko, C. Cheung, and J. Kavelaars. Artificial gravity: Head movements during short-radius centrifugation. *Acta Astronaut.* 49: 215–226 (2001).
17. Lackner, J.R., and P. DiZio. Human orientation and movement control in weightless and artificial gravity environments. *Exp. Brain Res.* 130: 2–26 (2000).
18. White, P.D., J.W. Nyberg, L.M. Finney, and W.J. White. *Influence of periodic centrifugation on cardiovascular functions of man during h\bed rest*, Report SM-48703. Douglas Aircraft Co. Inc. Santa Monica, CA, 1965.
19. Shulzhenko, E.B., I.F. Vil-Viliams, E.A. Aleksandrova, and K.I. Gogolev. Prophylactic effects of intermittent acceleration against physiological deconditioning in simulated weightlessness. *Life Sci. Space Res.* 17: 187–192 (1979).
20. Burton, R.R., and L.J. Meeker. Physiologic validation of a short-arm centrifuge for space application. *Aviat. Space Environ. Med.* 63: 476–481 (1992).
21. Yajima, K. K.I. Iwasaki, T. Sasaki, A. Miyamoto and K. Hirayanagi. Can daily centrifugation prevent the haematocrit increase elicited by 6-degree, head-down tilt? *Pflugers Archiv.* 441 Supplement (2–3): R95–97 (2000).
22. Iwasaki, K., K. Sasaki, K. Hirayanagi, and K. Yajima. Effects of repeated long duration + 2 G_z load on man's cardiovascular function. *Acta Astronaut.* 42 (1–8): 175–183 (1998).
23. Greenleaf, J.E., J.L. Chou, N.J. Stad, G.P.N. Leftheriotis, N. Arndt, C.G.R. Jackson, S.R. Simonson, and P.R. Barnes. Concomitant short-arm (1.9 m) + 2.2 G_x acceleration does not affect the isotonic exercise load-O₂ uptake relationship. *Aviation Space Environ. Med.* 70 (12): 1173–1182 (1999).
24. Vil-Viliams, I.F., and E.G. Shulzhenko. Functional state of the cardiovascular system under the combined effect of 28-day immersion, rotation on a short-arm centrifuge and exercise on a bicycle ergometer. *Kosm. Biol. Aviakosm. Med.* 14: 42–45 (1980).
25. Loret, B.J. Optimization of space vehicle design with respect to artificial gravity. *Aerosp. Med.* 34: 430–441 (1963).
26. Schultz, D.N., C.C. Rupp, G.A. Hajor, and J.M. Butler. A manned Mars artificial gravity vehicle. In *The Case for Mars III: Strategies for Exploration—General Interest and Overview*, C. Stoker (ed.) American Astronautical Society, pp. 325–352, 1989.

LAURENCE R. YOUNG

Massachusetts Institute of Technology
Cambridge, Massachusetts

ASTEROIDS

Asteroids are small members of the solar system in heliocentric orbits concentrated between Jupiter and Mars. Since most of them have orbits that are roughly similar to those of the planets (low inclination and eccentricity), they have sometimes been called minor planets, although this term is no longer in common use. More important is the distinction between asteroids and comets, where the primary difference is one of composition. Most asteroids are rocky objects, composed of the same sorts of materials as the inner planets. Comets, in contrast, contain a substantial quantity of water ice and other frozen volatiles

in addition to silicates and organic compounds. However, there are ambiguities in terminology. Comets that make frequent passes around the Sun may lose their volatiles and become indistinguishable from rocky asteroids. In addition, there are many volatile-rich objects being discovered in the outer solar system (beyond Neptune) that resist classification as either asteroids or comets. These are the Kuiper Belt Objects (KBOs), discussed in this Encyclopedia in the entry for Comets.

The most comprehensive references for asteroids are the large multiauthor textbooks published in the Space Science Series of the University of Arizona Press. See especially *Asteroids II* (1989) edited by R. Binzel and others, and *Hazards Due to Comets and Asteroids* (1994) edited by T. Gehrels. Soon to be published in the same series is *Asteroids III* (2002), edited by W. Bottke and others.

Discovery of Asteroids

Asteroids are too faint to be visible to the unaided eye, so their discovery belongs to the era of telescopic astronomy. On New Year's Day in 1801, Giuseppe Piazzi at Palermo Observatory found the first asteroid, which he named Ceres for the Roman patron goddess of Sicily. This faint object (now known as the largest asteroid) orbited the Sun at a distance of 2.8 astronomical units (AU). It was at first hailed as the "missing planet" in the large gap between the orbits of Mars and Jupiter. In the following few years, three more asteroids—Pallas, Juno, and Vesta—were found, also orbiting between Mars and Jupiter. These were even smaller than Ceres, although Vesta is slightly brighter due to its more reflective surface. Even combined, the masses of these four objects came nowhere near that of a real planet. Most of the asteroids are located in what is defined as the main asteroid belt, at distances from the Sun are between 2.2 and 3.3 AU, corresponding to orbital periods between 3.3 and 6.0 years.

The next asteroid was not discovered until 1845, but from then on, they were found regularly by visual observers who scanned the sky looking for them. By 1890, the total number had risen to 300. At that time, photographic patrols began, and the number of known objects rapidly increased and reached 1000 in 1923, 3000 in 1984, 5000 in 1990, and 20,000 in 2001. To be entered on the official list of asteroids, an object must be well enough observed to establish its orbit and permit its motion to be accurately calculated many years into the future. The responsibility for cataloging asteroids and approving new discoveries is assigned to the International Astronomical Union Minor Planet Center in Cambridge, Massachusetts (1). It is an interesting indication of growing interest that 198 years were required to find the first 10,000, but only two years were needed to find the second 10,000. Most recent discoveries are a by-product of the Spaceguard Survey, a concerted search for near-Earth Asteroids (NEAs) that will be discussed in detail at the end of this article.

In addition to numerical designations (e.g., 4 Vesta, 1000 Piazzia), which are given in chronological order of determination of an adequate orbit, most asteroids have names, usually suggested by the discoverer. Initially, these were the names of Greek and Roman goddesses, such as Ceres and Vesta, and later

expanded to include female names of any kind. When masculine names were applied, they were given the feminine Latin ending. More recently, the requirement of female gender has been dropped, and today asteroids are named for a bewildering variety of persons and places, famous or obscure. This article uses names rather than numbers to refer to specific asteroids.

Basic Asteroid Statistics

Our census of the larger asteroids is fairly complete by now, based primarily on ground-based surveys, complemented by infrared space observations, such as those from the Infrared Astronomy Satellite (IRAS) in 1983. It is likely that we have discovered all main-belt asteroids 25 km or more in diameter, and discovery should be more than 50% complete for diameters down to 10 km. Ceres, the largest, has a diameter just under 1000 km. The next largest asteroids are about half this size (see Table 1 for a listing of the dozen largest asteroids). The total mass of all of the asteroids amounts to only 1/2000 of the mass of Earth (less than 1/20 the mass of the Moon). Our knowledge is better for the closer asteroids in the inner part of the asteroid belt, and most of the larger undiscovered bodies are probably beyond 3 AU from the Sun.

There are many more small than large asteroids. An estimate of the relative numbers of objects of each size is interesting as a characterization of the asteroid population, and it is also closely related to the distribution of craters caused by the collision of asteroids with the planets and satellites in the inner solar system.

As a rule, many processes in nature, including those of fragmentation, result in approximately equal masses of material in each size range. Applying such a power law to the asteroids, we would find that there should be 1000 times more 10-km objects than 100-km ones, and a million more at 1 km than 100 km. In other words, the number of objects of a given diameter is inversely proportional to the cube of their diameter. However, measurements of the asteroids indicate that the numbers do not rise this fast as size declines but increase more nearly as the inverse square of the diameter, resulting in a distribution where most of the

Table 1. **The Largest Asteroids**

Name	Year of discovery	Distance from Sun, AU	Diameters, km	Class
Ceres	1801	2.77	940	C
Pallas	1802	2.77	540	C
Vesta	1807	2.36	510	V
Hygeia	1849	3.14	410	C
Ineramnia	1910	3.06	310	C
Davida	1903	3.18	310	C
Cybele	1861	3.43	280	C
Europa	1868	3.10	280	C
Sylvia	1866	3.48	275	C
Juno	1804	2.67	265	S
Psyche	1852	2.92	265	M
Patientia	1899	3.07	260	C

mass is in the larger objects. This is why we are relatively certain of the total mass of the asteroids, even without having counted all of the small ones.

From the observed distribution of sizes, we can estimate that there are more than 100,000 asteroids down to a diameter of 1 km. Although 100,000 sounds like a lot of objects, space in the asteroid belt is still empty. The belt asteroids occupy a very large volume, roughly doughnut shaped, about 100 million km thick and nearly 200 million km across. Typically the asteroids 1 km or larger are separated from each other by millions of kilometers. They pose no danger to passing spacecraft. In fact, it was challenging to locate asteroids near enough to the trajectory of outward-bound spacecraft (such as Galileo and Cassini) to allow close asteroid flybys on the way to Jupiter.

Physical and Chemical Properties

As seen through a telescope without special image compensation (adaptive optics), an individual asteroid is an unresolved starlike point. The word asteroid means starlike. Before about 1970, almost nothing was known about the physical nature of asteroids, and research was confined to discovering and charting orbits and determining rotational rates from observations of periodic variations in brightness. In the past 30 years, however, new observing techniques used with large telescopes have revealed a great deal about the physical and chemical nature of the asteroids (2). These astronomical observations have been supplemented by key studies of meteorites and by close-up spacecraft observations of a few asteroids, including Gaspra, Ida, Mathilde, and Eros—the latter involved both orbital and landed investigations.

Asteroid sizes and shapes are determined directly from imaging with modern adaptive optics or from the Hubble Space Telescope (although the resolution, even with the largest telescopes, leaves much to be desired). High-precision radar imaging is also a powerful tool if the object comes sufficiently close to Earth. Size can also be measured by timing the passage of an asteroid in front of a star. Because we know exactly how fast the asteroid is moving against the stellar background, measuring how long the star is obscured yields a chord length for the asteroid that can be accurate to a few kilometers. If timings of the same event made from different locations on Earth are combined, the profile of the asteroid can be derived. Unfortunately, however, suitable events are rare, and only a dozen asteroids have been measured successfully by this method.

Most asteroid sizes have been estimated indirectly from their visible or infrared brightness. Given only the apparent visible-light brightness of the object, we can roughly estimate its size by assuming a reflectivity or albedo that is characteristic of average asteroids. Such diameters are typically uncertain by a factor of 2, implying an order-of-magnitude uncertainty in mass. Much more accurate are reflectivities determined by combining visible-band measurement of reflected light with infrared-band measurement of emitted heat radiation. Such diameters are good to 10% or better, and they require no arbitrary assumptions about reflectivity.

It is clear that the asteroids have a variety of surface compositions, as discussed further later. This variety leads to a wide range of surface reflectivity.

The majority of the asteroids are very dark, roughly the brightness of charcoal. Other types can have reflectance as high as white terrestrial rocks. To make sense of this diversity of material, one must add information on the spectral reflectance of the asteroids. It is particularly useful to compare the asteroids with extraterrestrial samples that reach Earth as meteorites. A few meteorites come from the Moon or Mars, but the great majority of them are fragments from asteroids. Unfortunately, the chaotic dynamic processes that deliver meteorites to our planet do not include traceable return addresses. One of the major challenges of meteoritics is to connect the samples we have to their parent bodies (or class of parent bodies) among the asteroids.

The use of spectral data to characterize asteroids has yielded preliminary determinations of composition for approximately 1500 objects. These include a few asteroids that have metallic surfaces, presumably representing the surviving cores of objects that melted, differentiated chemically, and subsequently lost their stony crusts and mantles. Most, however, have rocky surfaces, that compare to the majority of meteorites, which are also rocky. Exact identifications are difficult, however, and usually we cannot specify the unique properties that identify an individual. With some notable exceptions, contemporary asteroid research, therefore, tends toward broad statistical studies rather than detailed investigation of particular objects. The exceptions are the handful of asteroids that have been visited by spacecraft or imaged at close range by radar, to be discussed further later.

Most of the well-observed asteroids fall into one of two classes based on their reflectivity (3). They are either very dark (reflecting only 3–5% of incident sunlight) or moderately bright (15–25% reflectivity). A similar distinction exists in their spectra. The dark asteroids are fairly neutral reflectors and do not have major absorption bands in the visible range to reveal their compositions, although some of them show spectral evidence of chemically bound water in the infrared. Most of the lighter asteroids are reddish and have the spectral signatures of common silicate minerals such as olivine and pyroxene. The dark gray asteroids have spectra similar to the carbonaceous meteorites, so they are called C-type asteroids. The lighter class is named the S-type, indicating silicate or stony composition. A third major group appears to be metallic (like the iron meteorites) and is called the M-type. There is also a variety of subclasses based on spectra and reflectivity, especially among the dark C-type objects. It is also increasingly clear that some process of “space weathering” alters the optical properties of surface materials; it partially masks identification with specific meteorite types on Earth and blurs the distinctions that might otherwise be seen among asteroids of different subgroups.

Using the classification of the asteroids, we can look at the distribution in space of the broad C, S, and M types. At the inner edge of the belt, the S asteroids predominate. Moving outward, the fraction of C-type objects increases steadily, and in the main asteroid belt as a whole, the dark, carbonaceous objects make up 75% of the population, compared to 15% S and 10% of M and other types.

Beyond the main belt, all asteroids are very dark, but their colors are redder than the belt objects, and they do not look like any known carbonaceous meteorite (4). Because these objects are not represented in our meteorite collections, scientists hesitate to commit themselves concerning their composition. It is

generally thought, however, that they are primitive objects and that a fragment from one of them would be classed as a carbonaceous meteorite, although of a kind different from those already encountered.

If the asteroids are still near the locations where they formed, we can use the distribution of asteroid types to map out the composition of the solar nebula, the original circumsolar assemblage of gas and dust from which the planetary system formed (5). Carbonaceous meteorites formed at lower temperatures than the other primitive stones, so we infer that the concentration of similarly composed C-type asteroids in the outer belt is consistent with their formation farther from the Sun, where the nebular temperatures were lower. It is also possible, however, that the asteroids formed elsewhere and were herded into their present positions by the gravity of Jupiter and the other planets. In that case, the C-type asteroids could have formed far beyond Jupiter and subsequently diffused inward to their present positions in the outer part of the asteroid belt. Similarly, the S-type asteroids near the inner edge of the belt could either have formed where we see them today, or they could have been gravitationally scattered to their present locations from still closer to the Sun. The solar nebula temperatures that we would deduce by applying these two alternative models are quite different. So far, however, we have not been able to settle on which model is preferred for the origin of the asteroids.

Orbits

The orbits of the belt asteroids are for the most part stable, their eccentricities are less than 0.3, and inclinations are below 20° . In the past, when presumably there were more asteroids, collisions may have been common, but by now the population has thinned to the point where each individual asteroid can expect to survive for billions of years between collisions. Still, with 100,000 objects 1 kilometer or more in size, a major collision somewhere in the belt is expected every few tens of thousands of years. Such collisions, as well as lesser cratering events, presumably yield some of the fragments that develop Earth-crossing orbits and eventually reach Earth as meteorites. In contrast, Earth-approaching NEAs have unstable orbits and typical dynamic lifetimes of only about 100 million years. Their numbers represent an equilibrium between inward scattering from the main belt and elimination either by colliding with the terrestrial planets or the Sun or by gravitational ejection from the solar system.

Given their history of collisions, there is no reason to expect that most asteroids are monoliths. Many may be rubble piles, consisting of loosely bound, low-density accumulations of debris that has reaccreted after a catastrophic disruption (6). In general, the energy required to disperse such debris completely is substantially greater than the energy needed to break up a target. One line of evidence for the existence of rubble piles comes from the highly elongated shapes of some small, rapidly spinning asteroids. These shapes are nearly what one might expect for an equipotential fluid and suggest such a reaccretion process. However, conclusive evidence of rubble piles awaited the first close-up spacecraft investigations, as recounted later.

The orbits of asteroids within the main belt are not evenly distributed. As shown in Fig. 1, some orbital periods are preferred, and others are nearly

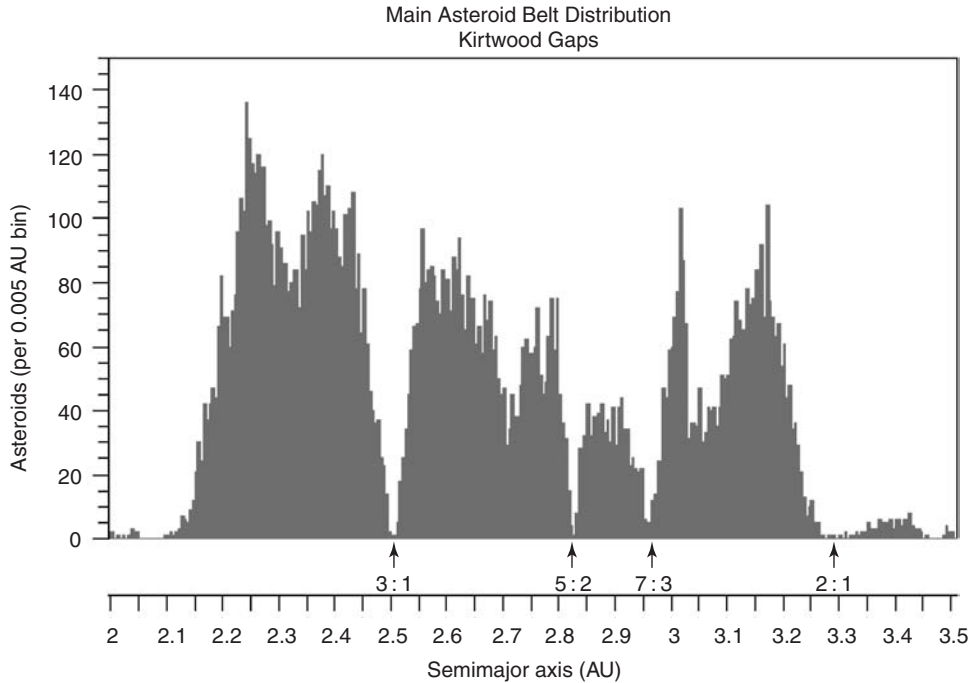


Figure 1. Histogram of orbital periods of the known asteroids. The deep minima are the Kirkwood gaps that correspond to periods that are in resonance with Jupiter (courtesy of Jet Propulsion Laboratory and NASA). Source: http://ssd.jpl.nasa.gov/a_histo.html. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

unpopulated. These unpopulated sections of the belt are resonance gaps, also known as the Kirkwood gaps for the nineteenth-century astronomer who discovered them. These gaps occur at orbital periods that correspond to resonances between these periods and the orbital period of Jupiter. Resonance takes place when the orbital period of one body is an exact fraction of the period of another. In this case, the underpopulated asteroid orbits correspond to periods that are one-half, one-third, one-quarter, etc., that of the 12-year orbital period of Jupiter. The Kirkwood gaps provide a clue to the origin of the asteroids or rather to the absence of a single large planet in the region between Mars and Jupiter. Presumably the dominant gravitational presence of Jupiter interrupted the accretionary process and dispersed the planetesimals in this part of the solar system. Most of the material ended up striking the inner planets or was ejected from the system, and only a small remnant remains in the asteroid belt today.

Asteroidal orbits display other patterns in addition to the resonance gaps. An asteroidal family is defined as a group of objects that have similar orbits that suggest a common origin. These were first identified by Kiyotsuga Hirayama early in the twentieth century. About half of the known belt asteroids are members of families, nearly 10% belong to just three: the Koronos, Eos, and Themis families. Although not clustered together in space at present, the members of an asteroid family were all at the same place at some undetermined time in the past. Members of the same family tend to have similar reflectivities and spectra.

Apparently, the family members are fragments of broken asteroids, shattered in some ancient collision, and still follow similar orbital paths. According to some estimates, almost all asteroids smaller than about 200 km in diameter were probably disrupted in earlier times, when the population of asteroids was larger. The families we see today may be remnants of the most recent of these inter-asteroidal collisions.

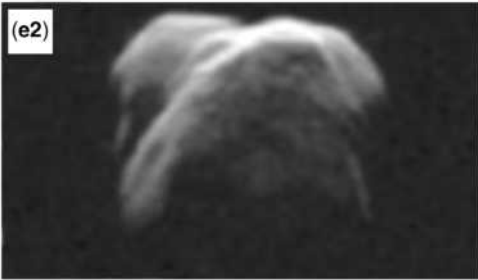
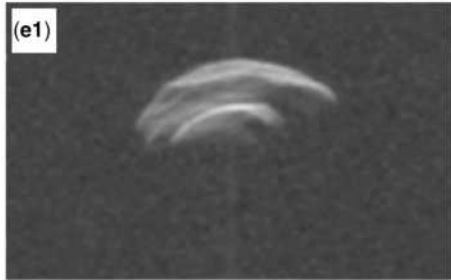
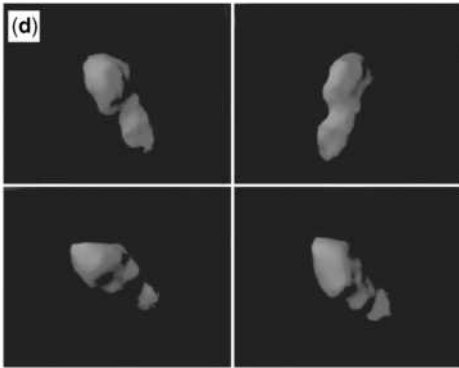
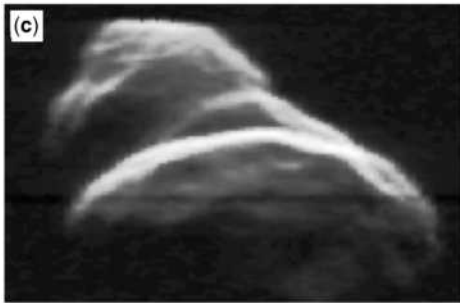
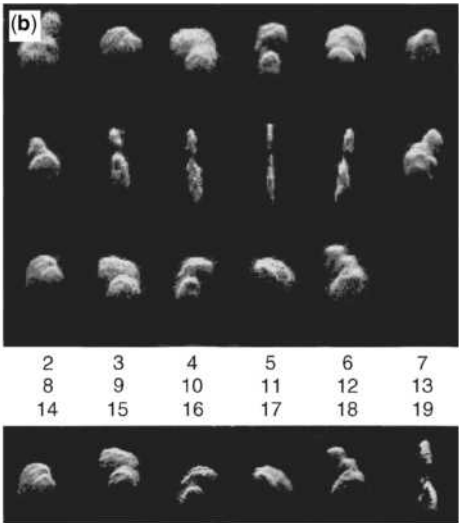
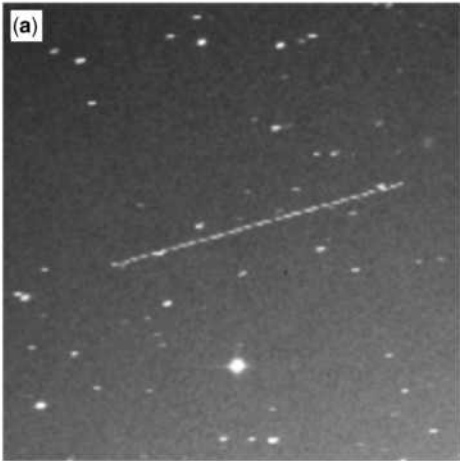
Asteroids up Close

Radar Studies. One of the most powerful tools for investigating asteroids is radar. There are two major planetary radar facilities, both of which were upgraded in the late 1990s. NASA operates the Goldstone (California) planetary radar facility as part of the Deep Space Net, and the 1000-foot Arecibo dish in Puerto Rico is operated by the National Astronomical and Ionospheric Center with NSF and NASA support. The two facilities are complementary—Arecibo has greater sensitivity, but Goldstone has greater sky coverage. Radar allows measuring range and velocity and permits us to define the rotational state precisely and to constrain the object's internal density distribution. In addition, radar astronomers used measurements of echo power in time delay (range) and Doppler frequency (radial velocity) to construct geologically detailed three-dimensional models that sometimes rival the resolution of spacecraft imaging systems (7).

By 2001, radar had detected more than 120 asteroids, whose sizes are as small as 30 m. These include large objects in the main belt as well as more than 80 of the smaller NEAs. One of the early radar contributions was to search for direct evidence of metallic surfaces for a few asteroids from their high microwave reflectivity. Observations of M asteroids Psyche and Kleopatra provide the best evidence linking the M class to metallic composition. However, these two asteroids have provided numerous surprises. In spite of its apparently metallic surface, Psyche has a density of only about 2 g/cm^3 , suggesting that its interior has extremely high porosity if composed of metal. Kleopatra is even stranger; it has a remarkable “dog-bone” shape that suggests reaccretion of material after a catastrophic impact. There is also evidence of a low-density surface of unconsolidated rubble on Kleopatra—again not what we would have expected by comparison with the lumps of iron-nickel in our meteorite collections (8).

The highest-resolution imaging has been achieved for asteroids that come very close to Earth. The largest of these is Toutatis, an elongated lumpy asteroid that provided early evidence that asteroids might not be monolithic (Fig. 2). At

Figure 2. Shape model from radar images of Toutatis. Analysis of the delay-Doppler imaging sequence established that Toutatis is in a non-principal-axis spin state, and accurate determination of the asteroid's rotation required inverting of the image sequence with a realistic physical model. These four views of the Toutatis computer model show shallow craters, linear ridges, and a deep topographic “neck” whose geologic origin is not known. It may have been sculpted by impacts into a single, coherent body, or this asteroid might actually consist of two separate objects that came together in a gentle collision. Toutatis is about 5 km long (images courtesy of Steve Ostro, NASA Jet Propulsion Lab and Scott Hudson, Washington State University). Source: four images labeled as “computer model” at http://echo.jpl.nasa.gov/asteroids/4179_Toutatis/toutatis.html.



5 km long, Toutatis is among the largest of the NEAs. Toutatis is also one of three asteroids found so far that are in slow, non-principal-axis spin states—perhaps evidence that they have received recent impacts (9). Among the interesting results of radar has been the discovery of three bifurcated objects (Castalia, Mithra, and Bacchus) that appear to be contact binaries. In several other cases, there is evidence of satellites orbiting asteroids. Satellites provide a way to calculate densities of the primary objects. Since the late 1990s, several asteroidal satellites (including the large C-type main-belt asteroids Eugenia and Antiope) have been discovered by using ground-based optical telescopes, and densities have also been measured for three of the asteroids visited by spacecraft. Most of the densities turn out to be surprisingly low (less than 2 g/cm^3), suggesting rather high interior porosity.

Spacecraft Flybys. Table 2 summarizes the four detailed spacecraft studies of asteroids. The main-belt asteroids Gaspra and Ida were flyby targets for Galileo on its way to Jupiter, and Mathilde was visited by the NEAR-Shoemaker spacecraft on its way to its primary target, Eros. The NEAR studies of Eros are discussed in the next subsection. Figure 3 illustrates these spacecraft targets on the same scale.

The Galileo flybys of two main-belt S-type asteroids revealed that both are highly irregular in shape, heavily cratered, and have only slight differences in color or reflectivity across their surfaces. Gaspra is undersaturated with craters, indicating a relatively young age (where age is the time since the last global-scale impact). In contrast, Ida is saturated with craters, and it appears to have a broken-up surface layer (a regolith) that is tens of meters thick (similar to that of the Moon). The discovery of a small satellite (Dactyl) in orbit around Ida permitted measuring its mass and density. The density is 2.6 g/cm^3 , similar to that of primitive rocks. Partly on this basis, it appears that these two S-type asteroids are probably coherent and are composed of materials similar to ordinary chondrite primitive meteorites. However, the spectral mismatch between these objects and known chondrites in our meteorite collections continued to baffle investigators after these two flybys. In addition, the presence of large families of grooves or lineaments on both asteroids suggested that they had global-scale cracks resulting from past impacts.

Mathilde was the first main-belt C-type asteroid to be examined at close range. NEAR-Shoemaker found a unique shape for this asteroid, dominated by several apparent craters whose diameters are greater than the radius of the asteroid. Such a configuration is not possible for a “normal” rocky target because the formation of the most recent of these craters would have been expected to destroy preexisting giant craters or perhaps even to disrupt the target entirely.

Table 2. Spacecraft Encounters with Asteroids

Asteroid	Class	Date	Dimensions, km	Density, g/cm^3	Best resolution, m
Gaspra	S	1991	$18 \times 11 \times 9$	—	50
Ida	S	1993	$60 \times 25 \times 19$	2.6	25
Mathilde	C	1997	$66 \times 48 \times 46$	1.3	160
Eros	S	2000	$31 \times 13 \times 13$	2.67	0.1

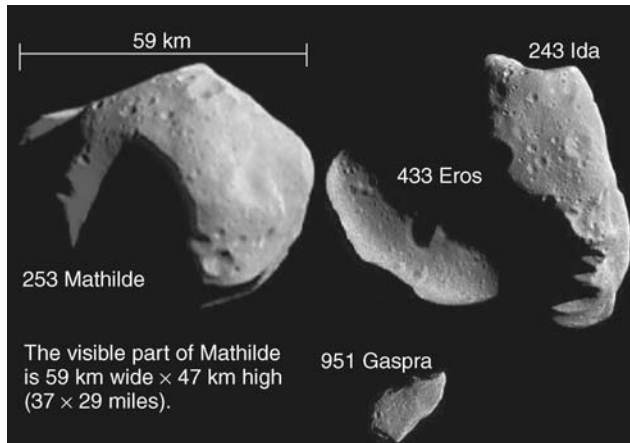


Figure 3. Family portrait of spacecraft images of asteroids Gaspra, Ida, Mathilde, and Eros, shown to scale. Gaspra and Ida (both main-belt asteroids) were imaged by the Galileo spacecraft. Mathilda (main-belt) and Eros (near-Earth) were imaged by NEAR-Shoemaker (Images courtesy of NASA, the Caltech Jet Propulsion Laboratory, and the JHU Applied Physics Laboratory.) Source: <http://junior.apk.net/~matto/comparison_mathild_ida_gaspra_eros.jpg>.

Only a “soft” target that has a less competent interior can absorb great shocks without internal disruption. This interpretation was reinforced by the measured density of 1.3 g/cm^3 , indicative of about 50% porosity. Thus, Mathilde became the first confirmed rubble-pile asteroid.

The NEAR-Shoemaker Mission to Eros. The most ambitious and successful spacecraft investigation of the asteroids was carried out by a small (Discovery-class) NASA spacecraft called the Near Earth Asteroid Rendezvous (NEAR) mission. It was further christened NEAR-Shoemaker in honor of Eugene Shoemaker, the father of asteroid geology. NEAR-Shoemaker missed its original rendezvous date with Eros in December 1998 due to a malfunction, but it recovered after one more trip around the Sun and finally arrived in February 2000. It achieved an initial high orbit, then gradually lowered its altitude during the next year, and studied Eros with a variety of instruments. The spacecraft obtained thousands of multispectral images and more than 10 million laser altimetry measurements, making Eros one of the best-mapped objects in the solar system (10,11).

After 1 year in orbit, NEAR-Shoemaker began a staged descent to the surface, taking pictures of ever-increasing resolution (Fig. 4). It landed on 12 February 2001 at an impact velocity of 1.6 m/s. Fortunately, the spacecraft was not damaged, even though it had not been designed for such a maneuver. Using its low-gain antenna, it continued radioing data from the surface for more than a week, providing the best measurements of elemental composition. The mission ended on 28 February when a command from Earth turned off the spacecraft (12).

The quantitative measurement of radioactivity from K, Th, and U, as well as gamma-ray lines of Fe, O, Si, and Mg, demonstrated that Eros has a primitive composition equivalent to the low-iron group of ordinary chondrite meteorites. Eros is a normal class-S asteroid, so this *in situ* result finally settled questions that had remained open for decades concerning the nature (primitive or



Figure 4. Close-up of near-Earth asteroid Eros as seen from the NEAR-Shoemaker cameras at a range of just 7 km. Most of the scene (about 350 meters across) is covered by rocks of all sizes and shapes, but the floors of some craters are smooth, suggesting accumulation of fine mobile material. The smallest visible features are about 1 meter across. (Image courtesy of NASA and the JHU Applied Physics Laboratory.) Source: <http://photojournal.jpl.nasa.gov/cgi-bin/PIAGenCatalogPage.pl?PIA03118>.

differentiated) of the S asteroids. The density of Eros (2.67 g/cm^3) is also generally consistent with this meteorite identification, although it still implies a substantial bulk porosity of about 25%. Evidently asteroids, like many terrestrial sediments, are consistently less dense than the individual rocks of which they are composed.

Long ridges seen in some of the images demonstrate that Eros is a consolidated and coherent body that has global-scale tectonics. As suspected for several other asteroids, Eros is a solid collisional fragment of a larger parent body (not a rubble pile), but it is also not a monolith because its interior has been heavily fractured. The surface is cratered, but there is a surprising deficiency of small craters, combined with an excess of boulders up to the 100-m size. There are actually more boulders than craters in the tens-of-meters sizes. Some measured slopes are greater than the angle of repose. Dark material has flowed down-slope, exposing underlying bright material. The effects of space weathering are evident in the different spectral reflectivity of exposures of differing age. Apparently Eros has a complex, mobile regolith, whose small-scale surface roughness is similar to that of lunar regolith (somewhat surprising because the gravity is so much less).

As noted in summer 2001 by MIT scientist Richard Binzel, “We’re getting to know asteroids as tangible objects, on the same scale and geologic sense that we know mountains on Earth.” And like terrestrial mountains, their interiors can be highly fragmented.

Trojan Asteroids

A particularly interesting group of dark, distant asteroids is orbitally associated with Jupiter. Although the gravitational attraction of this giant planet generally

makes nearby asteroidal orbits unstable, exceptions exist for objects of the same orbital period as Jupiter, while leading or trailing it by 60° . These two stable regions are called the leading and trailing Lagrangian points, named for the mathematician who demonstrated their existence in 1772. While he was mathematically examining the possible motions of three mutually gravitating bodies, Lagrange found two regions where a small object could occupy a stable orbit within the gravitational fields of two larger objects. If the larger objects are Jupiter and the Sun, a small object in one of the Lagrangian points occupies one corner of an equilateral triangle, and the Sun and Jupiter are at the other two points.

The regions of stability around the two Lagrangian points are quite large: each contains several hundred known asteroids. The first of these Lagrangian asteroids was named Hektor when it was discovered in 1907. All of them are named for the heroes of the *Iliad* who fought in the Trojan War, and collectively they are known as the Trojan asteroids. Their spectra are distinctive, suggesting that they represent a group of special, primitive objects that have been trapped in this region of space since the birth of Jupiter. If we could detect the fainter members of these Trojan clouds, we might find that the Trojan asteroids are nearly as numerous as those in the main asteroid belt.

Near-Earth Asteroids

Asteroid populations that can impact Earth are of special interest to us. They are generally referred to as Near-Earth Asteroids (NEAs) or Earth-crossing asteroids (ECAs). Because of their unstable, planet-approaching orbits, the NEAs have impacted the surfaces of the planets in the inner solar system (including Earth) and have influenced both geologic and biological evolution. There is reason to expect further impacts in the future, so the NEAs are a topic that has profound political and societal overtones. The impact hazard represents the intersection of asteroid science and public welfare and governmental policy (13).

It is highly improbable that a large (diameter > 1 km) NEA will hit the Earth within our lifetimes, but such an event is entirely possible. In the absence of specific information, such a catastrophe is equally likely at any time, including next year. Recognition that Earth (and Moon) are impacted by asteroids and comets is less than a century old, and it was not even securely proven that the prominent Meteor Crater (Arizona) was of impact origin until the work of Eugene Shoemaker in 1960. The fortunate fact that the atmosphere protects us from impacting bodies smaller than a few tens of meters in diameter (except for the rare iron meteorites) has the perhaps unfortunate consequence that we have almost no direct experience with cosmic impacts.

Tunguska and Meteor Crater. On the timescale of a human lifetime, the 1908 Tunguska impact in Siberia is the most notable. It was estimated (primarily from barographic and seismic records) that it had an explosive energy of ~ 15 megaton (TNT equivalent) when it disintegrated about 8 km above the ground. The impactor had the force of a large contemporary nuclear weapon. The explosion affected an unusually remote part of the world, and the first expedition to study Tunguska was delayed by two decades. At the time, before the existence of

an Earth-crossing asteroid population was recognized, it was naturally suggested that the culprit was a small comet. Other fringe-science explanations included the impact of a mini black hole and the crash of a UFO spacecraft. Not until the 1990s did numerical modeling of the entry physics clearly indicate that a comet (low-density, friable material) of this kinetic energy would disintegrate at very high altitudes and could not penetrate into the troposphere (14). Now we recognize that the event in Tunguska was simply the most recent example of an ongoing bombardment of Earth by NEAs.

A better known site of asteroidal impact is Meteor Crater (also called Barringer Crater) in northern Arizona. In this case, an iron asteroid about 40–50 meters in diameter struck about 50,000 years ago and formed a crater slightly more than 1 km in diameter. The energy of this impact was approximately the same as that of Tunguska (about 15 megaton), but because of the greater strength and density of the projectile, the explosion occurred at or below the surface, and a crater was formed.

Impacts and Extinctions. NEAs entered the scientific and popular mainstream in the 1980s when they were identified as the possible agents of biological mass extinctions. Alvarez and others (15) proposed that the dinosaur-killing KT mass extinction was due to an impact by a comet or asteroid, inferred from the chemical signature of extraterrestrial material in the boundary layer at the end of the Cretaceous. This bold hypothesis received general acceptance after the 200-km-diameter Chicxulub crater in Mexico (still among the largest craters identified on Earth) was discovered and it was dated exactly to the age of the KT extinction.

The most revolutionary insight of Alvarez and his colleagues was not that impacts take place on Earth (which was obvious), but that even small impacts (on a geological or astronomical scale) can severely damage the fragile terrestrial ecosystem. From the size of the Chicxulub crater, the energy of the KT impact is estimated at about 100 million megaton, and a consistent value of the size of the impactor (10–15 km in diameter) is derived from the observed extraterrestrial component in the boundary layer. Immediate effects of the impact included blast and the generation of a tsunami (because the impact occurred in a shallow sea). However, the primary agents of global stress appear to have been a short-lived firestorm from atmospheric heating of ejecta followed by a persistent (months to years) blackout due to particulates suspended in the stratosphere (16). Large land animals (such as the dinosaurs) were incinerated within a few minutes of the impact, and the marine ecosystem collapsed a few weeks later as a result of the global blackout. Fortunately, impacts of this size are exceedingly rare; they occur at average intervals of the order of a hundred million years. Today, there is no NEA comparable to the KT impactor that can hit Earth. However, we have no such assurance of immunity from smaller impacts.

Impacts from asteroids and comets have influenced the biological history of our planet in a variety of ways. It is widely thought that carbonaceous asteroids have been the dominant source of Earth's water and other volatiles, including many organic compounds required for originating life. At the same time, the impact environment of early Earth must have challenged the development of life and may have led to short episodes in which the oceans boiled away and the planet was sterilized. The phenomenon has been called the "impact frustration of life" (17). After the end of the heavy bombardment of Earth about 3.8 billion

years ago, impact catastrophes of this dimension were not possible. However, the Earth must have experienced dozens (or more) of impacts of the size of the KT event that punctuated biological evolution with occasional episodes of dramatic environmental stress. Impacts have been suspected in several other mass extinctions besides the KT, but in no other case is the evidence truly compelling. However, we know that these impacts have happened, and it is entirely plausible that they played a major role in biological evolution.

The Asteroid Impact Hazard

The average frequency of impacts by NEAs as a function of kinetic energy is illustrated in Fig. 5, adapted from a graph published in 1983 by Shoemaker (18). Comparison of this size–frequency distribution with the expected environmental damage caused by impacts of different energy leads to the conclusion (19) that the greatest risk is from large impacts, those that create a global ecological catastrophe. The threshold for global catastrophe is in the vicinity of 1 million megatons of energy, corresponding to an NEA whose diameter is about 2 km. Below this threshold, impacts create regional or local disasters, but the population (and social stability) of the planet are not threatened.

Although impacts below this million-megaton threshold are much more frequent, the total hazard from the sum of all such smaller impacts is less. Unlike more familiar natural hazards, the impact risk is primarily from extremely

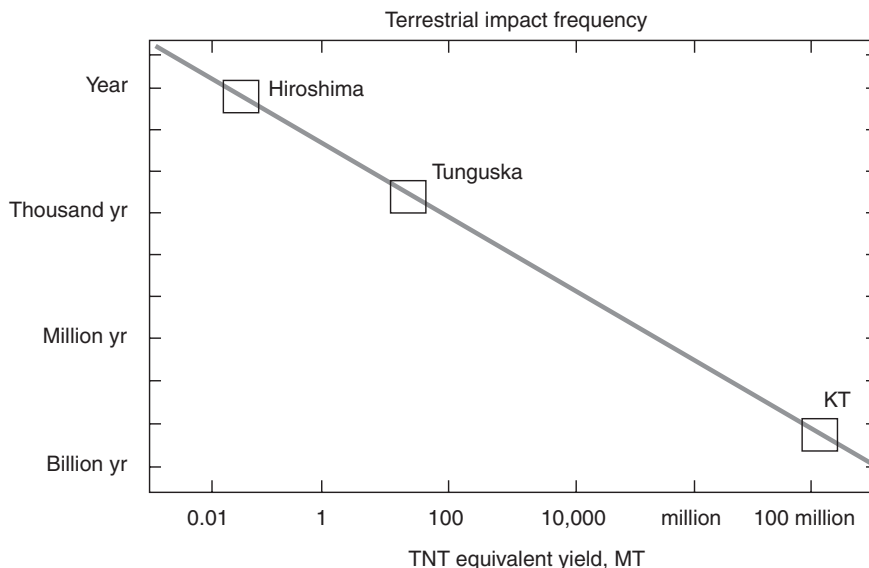


Figure 5. Plot of frequency of impacts on Earth vs. impact energy for near-Earth asteroids (NEAs). The power law is a long-term average derived primarily from lunar cratering and the current number and distribution of known NEAs. Shown plotted at their estimated energies are the Hiroshima nuclear bomb, the Tunguska impact of a small asteroid in Siberia (1908), and the KT impact that led to the extinction of the dinosaurs (65 million years ago). (Figure courtesy of David Morrison and NASA Ames Research Center.)

rare events—literally unprecedented in human history. The impact hazard represents the extreme case of a calamity of low probability but high consequences, including the possible end of civilization as we know it. It is logical to concentrate first on mitigating the risk from global catastrophes. Later, it may be desirable to extend mitigation efforts to smaller impacts that are much more likely to happen within our lifetimes, although they do not threaten society as a whole.

The preceding discussion treats impacts as if they are random statistical events, but they are in fact clearly deterministic. There either is or is not an NEA on a trajectory to collide with the Earth within, say, the next century. Any discussion of mitigation must recognize that these events can be predicted and even eliminated by deflecting a threatening NEA. The key requirement is adequate warning time. This is the philosophy behind the international “Spaceguard Survey” being carried out by ground-based optical telescopes equipped with state-of-the-art wide-field detectors and automated search capability (20). The NEAs are found as they repeatedly fly past Earth at typical distances of tens of millions of kilometers. If one of them should be on course for a future collision, it should be discovered decades (or more) in advance. The initial goal of Spaceguard is to discover and catalog at least 90% of all NEAs larger than 1 km in diameter within 10 years (by 2008). The focus is on NEAs of this size because 1 km is near the lower bound for an impact that can cause a global catastrophe. However, the observers continue to discover more NEAs below 1 km than above it, and over time the survey will extend completeness to smaller sizes.

The threat of impacts and the requirement to survey the NEAs have been recognized by the governments of the United States, the United Kingdom, Japan, and the Council of Europe, as well as by many *ad hoc* technical panels. The current Spaceguard Survey is being carried out by half a dozen observing teams primarily supported by NASA and the USAF. More than half the discoveries are being made by the Lincoln Lab/MIT group called LINEAR. As of the end of 2001, more than 500 of the estimated 1000 NEAs larger than 1 km have been found, and their orbits have been calculated. We can say with assurance that none of the discovered NEAs poses any threat on the timescale of a human lifetime, but of course we still cannot speak for the objects not yet discovered (21). How far the survey will be extended, and what plans will be developed for possible planetary protection against impacts, are questions for society as a whole, not just the small number of scientists who are currently studying NEAs.

Glossary

Asteroid. A small (diameter less than 1000 km) rocky or metallic solar system object in heliocentric orbit, generally of moderate or low orbital eccentricity and inclination. The majority are located in the asteroid belt between Mars and Jupiter. Sometimes called minor planet.

Asteroid Belt. The region in the solar system where most asteroids are found, between Mars and Jupiter. Specifically, objects in the main asteroid belt have orbital periods between 2.2 and 3.3 years.

Astronomical unit (AU). The mean distance of Earth from the Sun, approximately 150 million km.

Megaton (MT). Unit of energy equivalent to one million tons of TNT, 4.3×10^{15} joule.

Meteorite. Any extraterrestrial material that survives passage through the atmosphere and falls to Earth's surface. Most meteorites are fragments of asteroidal parent bodies.

Near-Earth Asteroid (NEA). An asteroid whose orbit brings it close to Earth (perihelion distance less than 1.3 AU) or, especially, an asteroid in an Earth-crossing orbit. Sometimes subdivided into Amor, Apollo, and Aten subgroups.

Regolith. The fragmented, dusty, porous, upper layer of material on a planetary surface; essentially the equivalent of Earth's soil for an object like the Moon that has little or no atmosphere or water.

BIBLIOGRAPHY

1. Bowell, E., N.S. Chernykh, and B.G. Marsden. Discovery and follow-up of asteroids. In R.P. Binzel, T. Gehrels, and M.S. Matthews (eds), *Asteroids II*. University of Arizona Press, Tuscon, 1989, pp. 21–38.
2. Binzel, R.P. An overview of the asteroids. In R.P. Binzel, T. Gehrels, and M.S. Matthews (eds), *Asteroids II*. University of Arizona Press, Tuscon, 1989, pp. 3–20.
3. Chapman, C.R., D. Morrison, and B. Zellner. Surface properties of asteroids: a synthesis of polarimetry, radiometry, and spectrophotometry. *Icarus* 25: 104–130 (1975).
4. Tholen, D.J., and M.A. Barucci. Asteroid taxonomy. In R.P. Binzel, T. Gehrels, and M.S. Matthews (eds), *Asteroids II*. University of Arizona Press, Tuscon, 1989, pp. 298–315.
5. Gradie, J.C., C.R. Chapman, and E.F. Tedesco. Distribution of taxonomic classes and the compositional structure of the asteroid belt. In R.P. Binzel, T. Gehrels, and M.S. Matthews (eds), *Asteroids II*. University of Arizona Press, Tuscon, 1989, pp. 316–335.
6. Asphaug, E., S.J. Ostro, R.S. Hudson, D.J. Scheeres, and W. Benz. Disruption of kilometre-sized asteroids by energetic collisions. *Nature* 393: 437–440 (1998).
7. Ostro, S.J., R.S. Hudson, L.A.M. Benner, J.D. Giorgini, C. Magri, J.-L. Margot, and M.C. Nolan. Asteroid radar astronomy. In W. Bottke et al. (eds), *Asteroids III*. University of Arizona Press, Tucson, 2002.
8. Ostro, S.J., R.S. Hudson, M.C. Nolan, J.-L. Margot, D.J. Scheeres, D.B. Campbell, C. Magri, J.D. Giorgini, and D.K. Yeomans. Radar observations of asteroid 216 Kleopatra. *Science* 288: 836–839 (2000).
9. Ostro, S.J., et al. Asteroid 4179 Toutatis: 1996 radar observations. *Icarus* 137: 122–139 (1999).
10. Veverka, J., et al. NEAR at Eros: Imaging and spectral results. *Science* 289: 1993–2228 (2000).
11. Zuber, M., et al. The Shape of 433 Eros from the NEAR-Shoemaker Laser Rangefinder. *Science* 289: 2097–2101 (2000).
12. Veverka, J., et al. The landing of the NEAR-Shoemaker spacecraft on asteroid 433 Eros. *Nature* 413: 390–393 (2001).
13. Morrison, D., A.W. Harris, G. Sommer, C.R. Chapman, and A. Carusi. Dealing with the impact hazard. In W. Bottke et al. (eds), *Asteroids III*. University of Arizona Press, Tucson, 2002.
14. Chyba, C.F., P.J. Thomas, and K.J. Zahnle. The 1908 Tunguska explosion: Atmospheric disruption of a stony asteroid. *Nature* 361: 40–44 (1993).
15. Alvarez, L., W. Alvarez, F. Asaro, and H.V. Michel. Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science* 208: 1095–1108 (1980).

16. Toon, O.B., K. Zahnle, D. Morrison, R.P. Turco, and C. Covey. Environmental perturbations caused by the impacts of asteroids and comets. *Rev. Geophys.* 35: 41–78 (1997).
17. Sleep, N.H., K.J. Zahnle, J.F. Kasting, and H.J. Morowitz. Annihilation of ecosystems by large asteroidal impacts on the early Earth. *Nature* 342: 139–142 (1989).
18. Shoemaker, E.M. Asteroid and comet bombardment of the Earth. *Ann. Rev. Earth Planetary Sci.* 11: 461–494 (1983).
19. Chapman, C.R., and D. Morrison. Impacts on the Earth by asteroids and comets: Assessing the hazard. *Nature* 367: 33–39 (1994).
20. Harris, A.W. Evaluation of ground-based optical surveys for near-Earth asteroids. *Planetary Space Sci.* 46: 283–290 (1998).
21. Binzel, R.P. The Torino impact hazard scale. *Planetary Space Sci.* 48: 297–303 (2000).

READING LIST

- Beatty, K. NEAR Falls for Eros. *Sky & Telescope* 102: 34–37 (2001).
- Binzel, R.P. A new century for asteroids. *Sky & Telescope* 102: 44–51 (2001).
- Binzel, R.P., T. Gehrels, and M.S. Matthews (eds). *Asteroids II*. University of Arizona Press, Tucson, 1994.
- Cowen, R. A rocky bicentennial: Asteroids come of age. *Science News* 160 (4): (2001).
- Gehrels, T. (ed.). *Hazards Due to Comets and Asteroids*. University of Arizona Press, Tucson, 1994.
- Lewis, J.S. *Rain of Iron and Ice: The Very Real Threat of Comet and Asteroid Bombardment*. Addison Wesley, New York, 1996.
- Morrison, D. Target Earth. *Astronomy* 30 (2): 46–51 (2002).
- Ostro, S.J. Radar observations of Earth-approaching asteroids. *Eng. Sci.* 60 (2): 24–31 (1997).
- Raup, D.M. *The Nemesis Affair: A Story of the Death of Dinosaurs and the Ways of Science*. Norton, New York, 1986.
- Steel, D. *Target Earth: The Search for Rogue Asteroids and Doomsday Comets*. Readers Digest, Pleasantville, 2000.

DAVID MORRISON
NASA Ames Research Center
Moffett Field, California

ASTROBIOLOGY

Astrobiology is a relatively new term that embraces the multidisciplinary study of the living Universe. It is the investigation of the origin, evolution, distribution, and destiny of life in the Universe. Astrobiology addresses some of the most profound questions of humankind: How did life begin? Are there other planets like Earth? What is our future as terrestrial life expands beyond the home planet? These questions are age-old. In the twenty-first century, however, advances in biological sciences, informatics, and space technology may make it possible for us to provide some answers.

Although the term had been used occasionally during previous decades as a synonym for “exobiology,” astrobiology in its present incarnation was proposed by NASA Associate Administrator for Space Science Wesley Huntress in 1995. NASA encouraged this new discipline by organizing workshops and technical meetings, establishing a NASA Astrobiology Institute, providing research funds to individual investigators, ensuring that astrobiology goals are incorporated in NASA flight missions, and initiating a program of public outreach and education. NASA’s role is derived from its history of studying the origin of life and searching for evidence of life on Mars or elsewhere in our solar system. Under the umbrella of astrobiology, these efforts are expanded to include the search for life within other planetary systems, as well as investigating the response of terrestrial life to global changes on Earth and to exposure to conditions in space and in other worlds. Astrobiology addresses our origins and also our aspirations to become a space-faring civilization.

Science Goals

The NASA Astrobiology road map (1) provides an initial description of the technical content of astrobiology. This road map was formulated through a series of workshops and discussions involving more than 400 people, primarily academic scientists who are interested in this new discipline. The road map represents a snapshot of a developing science and defines its content as perceived by scientists in 1998.

Astrobiology addresses three basic questions, which have been asked in some form for generations.

- How does life begin and evolve? (Where did we come from?)
- Does life exist elsewhere in the Universe? (Are we alone?)
- What is life’s future on Earth and beyond? (Where are we going in space?)

These are very general questions, and no one expects that definitive answers will be found easily. More specific is the analysis of astrobiology in terms of 10 long-term science goals.

1. Understand How Life Arose On Earth. Terrestrial life is the only form of life that we know, and it appears to have arisen from a common ancestor. How and where did this remarkable event occur? The question can be approached using historical, observational, and experimental investigations to understand the origin of life on our planet. We can describe the conditions on Earth when life began, use phylogenetic information to study our earliest ancestors, and also assess the possibility that life formed elsewhere and subsequently migrated to Earth.

2. Determine the General Principles Governing the Organization of Matter into Living Systems. To understand the full potential of life in the Universe, we must establish the general physical and chemical principles of life. We ask if terrestrial biochemistry and molecular biology are the only such phenomena that can support life. Having only one example, we do not know which properties of life are general and necessary and which are the result of specific

circumstances or historical accident. We seek these answers by pursuing laboratory experimental approaches and computational theoretical approaches.

3. Explore How Life Evolves on the Molecular, Organic, and Ecosystemic Levels. Life is a dynamic process of changes in energy and composition that occurs at all levels of assemblage from individual molecules to ecosystemic interactions. Modern genetic analysis, using novel laboratory and computational methods, allows new insights into the diversity of life and evolution at all levels. Complementary to such studies are investigations of the evolution of ecosystems consisting of many interdependent species, especially microbial communities.

4. Determine How the Terrestrial Biosphere Has Coevolved with Earth. Just as life evolves in response to changing environments, changing ecosystems alter Earth's environment. Astrobiologists seek to understand the diversity and distribution of our ancient ancestors by developing technology to read the record of life as captured in biomolecules and in rocks (fossils), to identify specific chemical interactions between the living components of Earth (its biosphere) and other planetary subsystems, and to trace the history of Earth's changing environment in response to external driving forces and to biological modifications.

5. Establish Limits For Life in Environments That Provide Analogs for Conditions in Other Worlds. Life is found on Earth anywhere liquid water is present, including such extreme environments as the interior of nuclear reactors, ice-covered Antarctic lakes, suboceanic hydrothermal vents, and deep subsurface rocks. To understand the possible environments for life in other worlds, we must investigate the full range of habitable environments on our own planet, for what they can tell us about the adaptability of life and also as analogs for conditions on other bodies in our solar system, such as Mars or Europa.

6. Determine What Makes a Planet Habitable and How Common These Worlds are in the Universe. Where should we look for extraterrestrial life? Based on our only example (life on Earth), liquid water is a requirement. Therefore, we must determine which sorts of planets are likely to have liquid water and how common they might be. Studying the processes of planet formation and surveying a representative sample of planetary systems will determine which planets are present and how they are distributed, essential knowledge for judging the frequency of habitable planets.

7. Determine How to Recognize the Signature of Life in Other Worlds. Astrobiologists need to learn to recognize extraterrestrial biospheres and to detect the signatures of extraterrestrial life. Within our own solar system, we must learn to recognize structural fossils or chemical traces of extinct life that may be found in extraterrestrial rocks or other samples (such as Martian meteorite ALH84001). To understand remotely sensed information from planets circling other stars, we should develop a catalog of possible spectral signatures of life. (See article on Extraterrestrial Life, Searching for in this Encyclopedia.)

8. Determine Whether There Is (or Once Was) Life Elsewhere in Our Solar System, Particularly on Mars and Europa. Exciting data have presented us with the possibility that at least two other worlds in our solar system have (or have had) liquid water present. On Mars, there is evidence for stable flowing water early in that planet's history. Both *in situ* investigations and the analysis of returned samples will be necessary to understand Mars' historical

climates and its potential for life. Because their surfaces are inhospitable, exploration of the subsurface probably offers the only credible opportunity to find extant life on either Mars or Europa.

9. Determine How Ecosystems Respond to Environmental Change on Timescales Relevant to Human Life on Earth. Research at the level of the whole biosphere is needed to examine the habitability of our planet over time in the face of both natural and human-induced environmental changes. To help ensure that continuing health of this planet and to understand the potential long-term habitability of other planets, we need to assess the role of rapid changes in the environment and develop our knowledge base to enable predictive models of environment–ecosystem interaction.

10. Understand the Response of Terrestrial Life to Conditions in Space or on Other Planets. All terrestrial life has developed in a one-gravity field, protected by Earth’s atmosphere and magnetic field. What happens when terrestrial life is moved off its home planet and into space or to the Moon or Mars, where the environment is very different from that of Earth? Can organisms and ecosystems adapt to a completely novel environment and live successfully for multiple generations? Are alternative strategies practical, such as bioengineering organisms for specific environments? The results from attempting to answer such questions will determine whether Earth’s life can expand its evolutionary trajectory beyond its place of origin.

Programmatic and Institutional Foundations

Astrobiology began as an effort within NASA to organize its space research programs around the theme of life in the Universe. It was given impetus by missions to Mars and Europa, by plans for telescopes in space to detect other planetary systems and measure the spectra of distant planets, and by the launch of the International Space Station and its planned suite of experimental facilities for life science. Within the NASA hierarchy, astrobiology has elements in the Space Science, Earth Science, and Human Exploration and Development of Space Enterprises. The lead management and coordination role was assigned to the Office of Space Science, and the lead NASA Center role was assigned to the Ames Research Center in California.

One of the early commitments to the development of astrobiology was the creation of a NASA Astrobiology Institute (NAI). This organization has the multiple objectives of encouraging commitments to astrobiology in the academic community, stimulating multidisciplinary research, and providing advice and technical input to NASA flight missions. Its member institutions are built around multidisciplinary research teams selected competitively. The central offices of the Astrobiology Institute are located at Ames Research Center, but the participating scientists (nearly 400 of them in 2001) remain employed in their own home institutions. Thus the NAI is a “virtual institute” or “collaboratory” in structure, using communications technology, together with an annual science meeting, postdoctoral fellows, and a number of cross-institutional “focus groups” to bind its geographically dispersed teams together. The first Director of the Astrobiology Institute is Nobel laureate Baruch Blumberg. Eleven

member institutions were selected in 1998: Harvard University, Marine Biological Laboratory at Woods Hole, Carnegie Institution of Washington, Pennsylvania State University, Arizona State University, Scripps Research Institute, University of California at Los Angeles, University of Colorado, NASA Ames Research Center, NASA Jet Propulsion Laboratory, and NASA Johnson Space Center. To these, the following were added in 2001: University of Rhode Island, Michigan State University, and University of Washington. Also associated with the NAI are the Center for Astrobiology in Madrid Spain and astrobiology teams in the United Kingdom and Australia. There is also a great opportunity for public access, as indicated by such popular web-sites as <astrobiology.arc.nasa.gov>, <nai.arc.nasa.gov>, and <www.astrobiology.com>.

The NASA astrobiology science goals bear a relationship to the search for extraterrestrial intelligence (SETI). Indeed, the detection of signals from an intelligent civilization on a distant planet would provide one of the most unambiguous signatures of extraterrestrial life (Goal 7 above). Historically, however, the SETI efforts have been separated from NASA since congressional action in 1993 terminated all NASA support for SETI programs. Intellectually, however, the two efforts represent complementary ways of addressing some of the same objectives.

Astrobiology is a science that has wide public appeal, as well as potential public concern. The search for life beyond Earth and the eventual expansion of terrestrial life to Mars or other planets in our solar system carry responsibility for protecting planetary ecosystems. Astrobiologists must ensure that these programs are carried out according to generally understood ethical and scientific principles. We will not endanger terrestrial life by introducing alien life-forms, and we will consider the broad ethical and cultural implications before we undertake to change the climate and surface conditions to make another world more hospitable to terrestrial life. Astrobiologists realize that their research has implications that are felt beyond the confines of the laboratory. As our understanding of living systems and the physical universe increases, we will confront the implications of this knowledge in more than just the scientific and technical realms. To understand the consequences will require multidisciplinary consideration of areas such as economics, environment, health, theology, ethics, quality of life, the sociopolitical realm, and education.

Summary

Astrobiology deals with a broad spectrum of disciplines, working together to use space technology to answer fundamental questions about life. Recent developments suggest that astrobiology is here to stay. But its success as a field will depend primarily on the quality of research carried out and on its contributions to space missions. If (or when) life is discovered on Mars or Europa, or the signature of life is detected in the light from an Earth-like planet circling another star, or commitments are made to human visits to other planets, then we can anticipate that astrobiology will hold center stage within space science.

Glossary

Astrobiology. The multidisciplinary study of the living universe, investigating the origin, evolution, distribution, and destiny of life in the Universe (term coined by Wesley Huntress in 1995). The term is somewhat broader than exobiology and includes studies of terrestrial life as it migrates into space.

Exobiology. The study of the origin, evolution, and distribution of life in the universe (term coined by Joshua Lederberg in 1960). In the United States, the term does not include studies of terrestrial life in space, but in Europe, “exobiology” sometimes takes on this broader meaning (which is essentially the same as “astrobiology”).

BIBLIOGRAPHY

1. Morrison, D. *Astrobiology* 1: 3–13 (2001).

READING LIST

- Goldsmith, D., and T. Owen. *The Search for Life in the Universe*, 3rd edition. University Science Books, Sausalito, CA, 2002.
- Boss, A. *Looking for Earths: The Race to Find New Solar Systems*. Wiley, New York, 1998.
- Woodward, C.E., J.M. Shull, and H.A. Thronson (eds), *Origins*. Astronomical Society of the Pacific Conference Series, San Francisco, 1998.
- Jakosky, B. *The Search for Life on Other Planets*. Cambridge University Press, New York, 1998.
- Goldsmith, D. *The Hunt for Life on Mars*. Dutton, New York, 1997. (New edition 1998).
- Davies, P.C.W. *The Fifth Miracle: The Search for the Origin and Meaning of Life*. Simon & Schuster, New York, 1999.
- Lemarchand, G., and K. Meech (eds), *Bioastronomy'99: A New Era in Bioastronomy*. Astronomical Society of the Pacific Conference Series, San Francisco, 1999.
- Lunine, J.I., and C.J. Lunine (illustrator). *Earth: Evolution of a Habitable World*. Cambridge University Press, New York, 1999.
- Ward, P., and D. Brownlee. *Rare Earth: Why Complex Life is Uncommon in the Universe*. Copernicus, New York, 2000.
- Dick, S.J. *Life on Other Worlds: The 20th-Century Extraterrestrial Life Debate*. Cambridge University Press, New York, 2000.

DAVID MORRISON
NASA Ames Research Center
Moffett Field, California

ASTRONAUTS AND THE PEOPLE WHO SELECTED THEM: A COMPENDIUM

The National Aeronautics and Space Administration's (NASA) Lyndon B. Johnson Space Center (JSC), located in Houston, Texas, has been responsible for conducting the astronaut recruiting and selection process. This includes

- establishing the astronaut staffing and selection requirements.
- issuing and releasing public announcements advertising the qualification requirements.
- appointing qualified members for rating panels and selection committees established for each recruiting campaign:
 - to review and perform the initial screening of applicants;
 - to conduct personal interviews of those applicants who satisfy initial screening requirements; and
 - to make final recommendations for selecting of applicants who qualify for the astronaut training program.

From the outset of human spaceflight, prospective candidates for the astronaut training program have been recruited on the basis of selection criteria dictated by the requirements of the missions to which they would be assigned. These mission requirements have grown in complexity; each follow-on program is responsible for advancing the state of the art of human space explorations. Such requirements have evolved from the pioneering effort in Project Mercury to those of far-reaching ramifications demonstrated in Project Apollo for landing humans on the Moon and more in the Space Shuttle and International Space Station Program for sustaining human life and supporting human productivity for flights of long duration.

In December 1958, President Dwight D. Eisenhower issued an edict limiting the pool of candidates for Project Mercury astronauts to military test pilots. Critics argued that this requirement excluded women, given the absence of female test pilots in the armed services to choose from and that the use of all military personnel in this highly visible position ran contrary to NASA's status as a civilian agency. However, Eisenhower's decision simplified and expedited NASA's selection process, eliminated the potential publicity blitz associated with an open call for applicants, and fit well with the highly technical and classified nature of certain aspects of the job. Screening boards for the Navy, Air Force, and Marines and a review by three aviation medicine specialists deemed 110 test pilots suitably qualified for further evaluation by the NASA Space Task Group, the organization charged with putting America's first human in space. Robert R. Gilruth, the Space Task Group Director, delegated all authority for selecting the Mercury astronauts to his deputy, Charles J. Donlan, and his handpicked board. Though rigorous medical and psychological tests provided a wealth of information on each potential astronaut, this first selection committee also weighed its impressions gained through personal interviews with each candidate. The committee regarded individual initiative and complementary technical expertise within the group as the deciding factors in March 1959 when they chose the seven test pilots who would be America's first men in space. In the following month, after approval by Gilruth and NASA management, the space agency publicly announced the names of the Mercury Seven astronauts.

The Original Seven

The initial astronaut selection began in January 1959, before human spaceflight operations began, when NASA asked the military services to screen personnel

records for prospective candidates who met the qualifications outlined in Table 1. This preliminary review produced 508 military test pilots, of whom 110 satisfied all of the basic requirements. A subsequent examination of military and medical records reduced the total eligible to 69, all of whom were invited to Washington for a briefing on Project Mercury and a personal interview. It was on the basis of these interviews, that 32 men were chosen to undergo detailed physical examination and stress testing; however, it became apparent during this phase that all of the candidates surpassed the established medical standards. Consequently, final screening concentrated on individual engineering and operational performance, and the seven best technically qualified were selected in April 1959 for the Mercury Program.

Note: All ranks for military and uniformed services officers referenced indicate the rank held when the individual was selected for the astronaut program.

M. Scott Carpenter, Lt., USN. Born May 1, 1925. B.S. in Aeronautical Engineering from the University of Colorado.

L. Gordon Cooper, Capt., USAF. Born March 6, 1927. B.S. in Aeronautical Engineering from the Air Force Institute of Technology.

John H. Glenn, Jr., Lt. Col., USMC. Born July 18, 1921. B.S. in Engineering from Muskingum College.

Virgil I. Grissom, Capt., USAF. Born April 3, 1926. B.S. in Mechanical Engineering from Purdue University.

Walter M. Schirra, Jr., Lt. Comdr., USN. Born March 12, 1923. B.S. from the U.S. Naval Academy.

Alan B. Shepard, Lt. Comdr., USN. Born November 18, 1923. B.S. from the U.S. Naval Academy.

Donald K. Slayton, Capt., USAF. Born March 1, 1924. B.S. in Aeronautical Engineering from the University of Minnesota.

An ad hoc selection committee was appointed for each recruiting effort conducted in the search for candidates who qualified for the astronaut training program. Participating in the selection of the original seven were Charles J. Donlan, NASA Space Task Group; A. O. Gamble, NASA Headquarters; Robert R. Gilruth, NASA Space Task Group; and Warren J. North, NASA Headquarters.

Group 2

In April 1962, an announcement was issued from Houston to recruit a second group of astronauts to train for the Gemini and Apollo Programs. Minimum qualification standards were published and disseminated to aircraft companies, government agencies, military services, the Society of Experimental Test Pilots, and the news media (see Table 1). A total of 250 applications was received from civilian and military sources. Each candidate who satisfied the five basic standards was asked to complete a variety of forms describing academic credentials and flight and work experience in detail. Each was also asked to submit to a thorough physical/medical examination and to forward the results to the Johnson Space

Table 1. **Astronaut Selection History^a 1958 to 2000**

Group	Date	Number selected	Max age	Max height	Min degree level	Jet pilot	TPS grad	Flying time	Experience	Outside help
1	Apr. 59	7	39	5'-10"	B.S.	Yes	Yes	1500		
2	Sept. 62	9	34	6'-0"	B.S.	Yes	Yes	1500		
3	Oct. 63	14	32	6'-0"	B.S.	Yes	Optional	1000		
4	June 65	6	34	6'-0"	Ph.D.	No	No	No		Nas ^b
5	Apr. 66	19	34	6'-0"	B.S.	Yes	Optional	1000		
6	Aug. 67	11	36	6'-0"	Ph.D.	No	No	No		Nas ^b
7	Aug. 69	7	35	Transferred to NASA upon cancellation of USAF MOL program						
8 thru 18	78 to '00									
Pilot			N/A	64 TO 76 in	B.S.	Yes	Preferable	1000		
Mssn spec			N/A	58.5 TO 76 in	B.S.	No			3 years	

^aAll selections required U.S. citizenship.

^bNAS is the National Academy of Science.

Center (then the Manned Spacecraft Center) in Houston, Texas. In June 1962, a preliminary selection committee reviewed this additional information submitted by the individual candidates and selected 32 of the most qualified applicants to participate in further examinations, tests, and personal interviews. Nine pilot astronauts comprised the group finally selected in September 1962:

Neil A. Armstrong, civilian. Born August 5, 1930. B.S. in Aeronautical Engineering from Purdue University and attended graduate school at the University of Southern California.

Frank Borman, Maj., USAF. Born March 14, 1928. B.S. in Aeronautical Engineering from the U.S. Military Academy and M.S. in Aeronautical Engineering from the California Institute of Technology.

Charles Conrad, Jr., Lt., USN. Born June 2, 1930. B.S. in Aeronautical Engineering from Princeton University.

James A. Lovell, Jr., Lt. Comdr., USN. Born March 25, 1928. B.S. from the U.S. Naval Academy.

James A. McDivitt, Capt., USAF. Born June 10, 1929. B.S. in Aeronautical Engineering from the University of Michigan.

Elliot M. See, Jr., civilian. Born July 23, 1927. B.S. from the U.S. Merchant Marine Academy and M.S. in Engineering from the University of California at Los Angeles.

Thomas P. Stafford, Capt., USAF. Born September 17, 1930. B.S. from the U.S. Naval Academy.

Edward H. White II, Capt., USAF. Born November 14, 1930. B.S. from the U.S. Military Academy and M.S. in Aeronautical Engineering from the University of Michigan.

John W. Young, Lt. Comdr., USN. Born September 24, 1930. B.S. in Aeronautical Engineering from Georgia Institute of Technology.

The committee for the selection of Group 2 included Warren J. North, NASA JSC; Alan B. Shepard, Jr., NASA JSC; and Donald K. Slayton, NASA JSC.

Group 3

A third call for applications for the astronaut process was issued in June 1963. For this group, the requirement for test pilot school was optional (see Table 1), and the required jet pilot time was reduced to 1000 hours. With this decrease in actual flying requirements, increased emphasis was given to academic areas. A total of 720 applications was received—228 from civilians and 492 from military personnel. Of the 490 certified eligibles, 136 were referred for final screening by the NASA selection board. Fourteen new astronauts were named in October 1963. Two civilians, seven Air Force pilots, four Navy aviators, and one Marine Corps aviator comprised this third group of trainees:

Edwin E. Aldrin, Jr., Maj., USAF. Born January 20, 1930. B.S. from the U.S. Military Academy and Sc.D. in Astronautics from Massachusetts Institute of Technology.

William A. Anders, Capt., USAF. Born October 17, 1933. B.S. from U.S. Naval Academy and M.S. in Nuclear Engineering from the Air Force Institute of Technology.

Charles A. Bassett II, Capt., USAF. Born December 30, 1931. B.S. in Electrical Engineering from Texas Technological University.

Alan L. Bean, Lt., USN. Born March 15, 1932. B.S. in Aeronautical Engineering from the University of Texas.

Eugene A. Cernan, Lt., USN. Born March 14, 1934. B.S. in Electrical Engineering from Purdue University and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Roger B. Chaffee, Lt., USN. Born February 15, 1935. B.S. in Aeronautical Engineering from Purdue University.

Michael Collins, Capt., USAF. Born October 31, 1930. B.S. from the U.S. Military Academy.

R. Walter Cunningham, civilian. Born March 16, 1932. B.A. and M.A. in Physics from the University of California at Los Angeles.

Donn F. Eisele, Capt., USAF. Born June 23, 1930. B.S. from the U.S. Naval Academy and M.S. in Astronautics from the Air Force Institute of Technology.

Theodore C. Freeman, Capt., USAF. Born February 18, 1930. B.S. from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the University of Michigan.

Richard F. Gordon, Lt. Comdr., USN. Born October 5, 1929. B.S. in Chemistry from the University of Washington.

Russell L. Schweickart, civilian. Born October 25, 1935. B.S. in Aeronautical Engineering and M.S. in Aeronautics and Astronautics from Massachusetts Institute of Technology.

David R. Scott, Capt., USAF. Born June 6, 1932. B.S. from the U.S. Military Academy, and M.S. and M.E. in Aeronautics and Astronautics from Massachusetts Institute of Technology.

Clifton C. Williams, Jr., Capt., USMC. Born September 26, 1932. B.S. in Mechanical Engineering from Auburn University.

The selection committee members for Group 3 were John H. Glenn, Jr., NASA JSC; Warren J. North, NASA JSC; Walter M. Schirra, Jr., NASA JSC; Alan B. Shepard, Jr., NASA JSC; and Donald K. Slayton, NASA JSC.

Group 4 (Scientist-Astronauts)

NASA began recruiting for its first group of scientist-astronauts in October 1964. For this group, flying status was desirable but not a mandatory prerequisite for selection (see Table 1). However, each of those selected was required to pass a Class I military flight physical examination before acceptance to the training program. Emphasis was on graduate work in the natural sciences, such as physics, medicine, engineering, or comparable occupational experience. A total

of 1492 letters of interest was received in Houston by January 1965. Some were informal inquiries, but 909 were formal applications. Of the latter, 424 qualified under the minimum criteria established and were forwarded to the National Academy of Sciences in Washington DC for evaluation. The Academy evaluated these applications for conformity with scientific criteria developed cooperatively with the NASA Office of Space and Applications and recommended 16 candidates for final consideration. These 16 applicants underwent thorough physical examinations and stress testing, and six were selected in June 1965 for training as scientist-astronauts. One geologist, two physicians, and three physicists comprised the group, and two in the group were qualified jet pilots. Those without jet pilot experience underwent one year's flight training before entering the regular astronaut training program. The following were selected:

Owen K. Garriott, civilian. Born November 22, 1930. B.S. in Electrical Engineering from the University of Oklahoma and M.S. and Ph.D. in Electrical Engineering from Stanford University.

Edward G. Gibson, civilian. Born November 8, 1936. B.S. in Engineering from the University of Rochester and M.S. in Engineering and Ph.D. in Engineering and Physics from California Institute of Technology.

Duane E. Graveline, civilian. Born March 2, 1931. Ph.D./M.D. from the University of Vermont and M.S. in Public Health from Walter Reed Army Medical Center.

Joseph P. Kerwin, Lt. Comdr., USN. Born February 19, 1932. B.A. in Philosophy from the College of the Holy Cross and M.D. from Northwestern University Medical School.

F. Curtis Michel, civilian. Born June 5, 1934. B.S. and Ph.D. in Physics from California Institute of Technology.

Harrison H. Schmitt, civilian. Born July 3, 1935. B.S. in Science from California Institute of Technology and Ph.D. in Geology from Harvard University.

The selection committee for Group 4 consisted of the following members from the National Academy of Sciences Board: Dr. Allan H. Brown, Department of Biology, Joseph Leidy Laboratory of Biology, University of Pennsylvania; Professor L.D. Carlson, Department of Physiology, University of California Medical School; Professor Frederick L. Ferris, Jr., Educational Services, Inc.; Dr. Thomas Gold, Chairman, Astronomy Department, Director, Center for Radio, Physics and Space Research, Cornell University; Dr. H. Keffer Hartline, Rockefeller University; Dr. Clifford T. Morgan, Department of Psychology, University of California; Dr. Eugene M. Shoemaker, Astrogeology Branch, U.S. Geological Survey; Dr. Robert Speed, Department of Geology, Northwestern University; and Professor Aaron C. Waters, Department of Geology, University of California.

The NASA Board consisted of the following individuals: Charles A. Berry, M.D., NASA JSC; John F. Clark, NASA Goddard Space Flight Center (GSFC); Maxime A. Faget, NASA JSC; Warren J. North, NASA JSC; Alan B. Shepard, Jr., NASA JSC; and Donald K. Slayton, NASA JSC.

Group 5

The Johnson Space Center launched its fifth recruiting drive in September 1965. Eligibility requirements were basically the same as those used in selecting the third group of astronaut trainees (see Table 1). A total of 510 applications was received, of which 158 (100 military and 58 civilians) met basic requirements. The previously established screening procedures were followed, yielding 19 pilot-astronauts who were selected in April 1966. Selectees were:

Vance D. Brand, civilian. Born May 9, 1931. B.S. in Business and Aeronautical Engineering from the University of Colorado and MBA from the University of California at Los Angeles.

John S. Bull, Lt., USN. Born September 25, 1934. B.S. in Mechanical Engineering from Rice University.

Gerald P. Carr, Maj., USMC. Born August 22, 1932. B.S. in Mechanical Engineering from the University of Southern California, B.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School, and M.S. in Aeronautical Engineering from Princeton University.

Charles M. Duke, Jr., Capt., USAF. Born October 3, 1935. B.S. in Naval Sciences from the U.S. Naval Academy and M.S. in Aeronautics and Astronautics from Massachusetts Institute of Technology.

Joe H. Engle, Capt., USAF. Born August 26, 1932. B.S. in Aeronautical Engineering from the University of Kansas.

Ronald E. Evans, Lt. Comdr., USN. Born November 10, 1933. B.S. in Electrical Engineering from the University of Kansas and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Edward G. Givens, Jr., Maj., USAF. Born January 5, 1930. B.S. in Naval Sciences from the U.S. Naval Academy.

Fred W. Haise, Jr., civilian. Born November 14, 1933. B.S. in Aeronautical Engineering from the University of Oklahoma.

James B. Irwin, Maj., USAF. Born March 17, 1930. B.S. in Naval Sciences from the U.S. Naval Academy and M.S. in Aeronautical Engineering and Instrumentation Engineering from the University of Michigan.

Don L. Lind, civilian. Born May 18, 1930. B.S. in Physics from the University of Utah and Ph.D. in High Energy Nuclear Physics from the University of California at Berkeley.

Jack R. Lousma, Capt., USMC. Born February 29, 1936. B.S. in Aeronautical Engineering from the University of Michigan and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Thomas K. Mattingly II, Lt., USN. Born March 17, 1936. B.S. in Aeronautical Engineering from Auburn University.

Bruce McCandless II, Lt., USN. Born June 8, 1937. B.S. in Naval Sciences from the U.S. Naval Academy and M.S. in Electrical Engineering from Stanford University.

Edgar D. Mitchell, Comdr., USN. Born September 17, 1930. B.S. in Industrial Management from Carnegie Institute of Technology, B.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School, and Sc.D. in Aeronautics and Astronautics from Massachusetts Institute of Technology.

William R. Pogue, Maj., USAF. Born January 23, 1930. B.S. in Education from Oklahoma Baptist University and M.S. in Mathematics from Oklahoma State University.

Stuart A. Roosa, Capt., USAF. Born August 16, 1933. B.S. in Aeronautical Engineering from the University of Colorado.

John L. Swigert, Jr., civilian. Born August 30, 1931. B.S. in Mechanical Engineering from the University of Colorado, M.S. in Aerospace Science from Rensselaer Polytechnic Institute, and M.S. in Business Administration from the University of Hartford.

Paul J. Weitz, Lt. Comdr., USN. Born July 25, 1932. B.S. in Aeronautical Engineering from Pennsylvania State University and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Alfred M. Worden, Capt., USAF. Born February 7, 1932. B.S. from the U.S. Military Academy and M.S. in Aeronautical/Astronautical Engineering and Instrumentation Engineering from the University of Michigan.

The selection committee members for Group 5 were Charles Conrad, Jr., NASA JSC; L. Gordon Cooper, Jr., NASA JSC; Virgil I. Grissom, NASA JSC; Warren J. North, NASA JSC; Donald K. Slayton, NASA JSC; Clifton C. Williams, Jr., NASA JSC; and John W. Young, NASA JSC.

Group 6 (Scientist-Astronauts)

In September 1966, NASA requested the National Academy of Sciences to nominate a second group of scientist-astronauts. NASA encouraged the Academy to seek experienced scientists of exceptional ability “to conduct scientific experiments in manned orbiting satellites and to observe and investigate the lunar surface and circumterrestrial space.” The Academy then issued its announcement stating: “The quality most needed by a scientist serving as an astronaut might be summed up by the single word “perspicacity.” The task requires an exceptionally astute and imaginative observer but also one whose observations are accurate and impartial. He must, from among the thousands of items he might observe, quickly pick out those that are significant, spot the anomalies, and investigate them. He must discriminate fine detail and subtle insight into a general pattern, and select and devise key observations to test working hypotheses. The selection criteria and procedures were comparable to those used in choosing the first group of scientist-astronauts (see Table 1). Nine-hundred applicants responded, but the Academy recommended only 69 for NASA’s final consideration. Of this number, 11 were chosen:

Joseph P. Allen IV, civilian. Born June 27, 1937. B.A. in Math-Physics from De Pauw University and M.S. and Ph.D. in Physics from Yale University.

- Philip K. Chapman, civilian. Born March 5, 1935. B.S. in Physics and Mathematics from Sydney University; M.S. in Aeronautics and Astronautics and Sc.D. in Instrumentation from Massachusetts Institute of Technology.
- Anthony W. England, civilian. Born May 15, 1942. B.S. and M.S. in Geology and Physics from Massachusetts Institute of Technology.
- Karl G. Henize, civilian. Born October 17, 1926. B.A. in Mathematics and M.A. in Astronomy from the University of Virginia and Ph.D. in Astronomy from the University of Michigan.
- Donald L. Holmquest, civilian. Born April 7, 1939. B.S. in Electrical Engineering from Southern Methodist University and doctorates in Medicine and Physiology from Baylor University.
- William B. Lenoir, civilian. Born March 14, 1939. B.S., M.S., and Ph.D. in Electrical Engineering from Massachusetts Institute of Technology.
- John A. Llewellyn, civilian. Born April 22, 1933. B.S. and Ph.D. in Chemistry from the University College of Cardiff.
- F. Story Musgrave, civilian. Born August 19, 1935. B.S. in Mathematics and Statistics from Syracuse University, M.B.A. in Operations Analysis and Computer Programming from the University of California at Los Angeles, B.A. in Chemistry from Marietta College, M.D. from Columbia University, and M.S. in Biophysics from the University of Kentucky.
- Brian T. O'Leary, civilian. Born January 27, 1940. B.A. in Physics from Williams College, M.A. in Astronomy from Georgetown University, and Ph.D. in Astronomy from the University of California at Berkeley.
- Robert A. R. Parker, civilian. Born December 14, 1936. B.A. in Astronomy and Physics from Amherst College and Ph.D. in Astronomy from California Institute of Technology.
- William E. Thornton, civilian. Born April 14, 1929. B.S. in Physics and M.D. from the University of North Carolina.

The Group 6 selection committee members from the National Academy of Sciences Board were Dr. Allan H. Brown, Department of Biology, Joseph Leidy Laboratory of Biology, University of Pennsylvania; Professor L.D. Carlson, Department of Physiology, University of California Medical School; Dr. Arthur B. Dubois, Division of Graduate Medicine, Department of Physiology, University of Pennsylvania; and Dr. H. Keffer Hartline, Rockefeller University.

The Life Sciences Subpanel members were Dr. George V. LeRoy, Medical Director, Metropolitan Hospital, Detroit, Michigan; Dr. Clifford T. Morgan, Department of Psychology, University of California; and Dr. Norton Nelson, Provost, University Heights Center, New York University.

The Physical Sciences Subpanel members were Dr. Edward W. Cannon, Chief, Applied Mathematics Division, National Bureau of Standards; Professor Frederick L. Ferris, Jr., Educational Services, Inc.; Dr. Harry H. Hess, Department of Geology, Princeton University; Dr. John D. Hoffmann, Chief, Polymers Division, National Bureau of Standards; Dr. Phillip Mange, Naval Research Laboratory; Dr. Eberhardt Rechtin, Assistant Director for Tracking and Data Acquisition, NASA JPL; Dr. Eugene M. Shoemaker, Astrogeology Branch, U.S. Geological Survey;

Dr. Shirleigh Silverman, Associate Director of Academics Liaison, National Bureau of Standards; Professor Philip N. Slater, Research Professor, Stewart Observatories, University of Arizona; Dr. Robert Speed, Department of Geology, Northwestern University; Professor Edward C. Stevenson, Professor of Electrical Engineering, University of Virginia; Professor Aaron C. Waters, Department of Geology, University of California; and Dr. Arthur H. Waynick, Director, Ionosphere Research Laboratory, Pennsylvania State University.

The NASA Board members were Charles A. Berry, M.D., NASA JSC; Maxime A. Faget, NASA JSC; Owen K. Garriott, Ph.D., NASA JSC; Wilmot N. Hess, NASA JSC; Alan B. Shepard, Jr., NASA JSC; Donald K. Slayton, NASA JSC; and Robert F. Thompson, NASA JSC.

Group 7

This group of seven pilot astronauts, transferred to NASA from the USAF Manned Orbiting Laboratory (MOL) Program when it was cancelled in August 1969. Although there were 13 in the MOL contingent, NASA absorbed only those under the age of 35:

Karol J. Bobko, Maj., USAF. Born December 23, 1937. B.S. from the U.S. Air Force Academy and M.S. in Aerospace Engineering from the University of Southern California.

Robert L. Crippen, Lt. Comdr., USN. Born September 11, 1937. B.S. in Aerospace Engineering from the University of Texas.

C. Gordon Fullerton, Maj., USAF. Born October 11, 1936. B.S. and M.S. in Mechanical Engineering from California Institute of Technology.

Henry W. Hartsfield, Maj., USAF. Born November 21, 1933. B.S. in Physics from Auburn University.

Robert F. Overmyer, Maj., USMC. Born July 14, 1936. B.S. in Physics from Baldwin-Wallace College and M.S. in Aeronautics from the U.S. Naval Postgraduate School.

Donald H. Peterson, Maj., USAF. Born October 22, 1933. B.S. from the U.S. Military Academy and M.S. in Nuclear Engineering from the Air Force Institute of Technology.

Richard H. Truly, Lt. Comdr., USN. Born November 12, 1937. B.S. in Aeronautical Engineering from Georgia Institute of Technology.

Group 8—Space Shuttle Astronauts

In 1978, NASA selected 35 astronaut candidates as the first group to support the Space Shuttle program. One of the 15 pilots was African-American. Six females, two African-Americans, and one Asian-Pacific Islander were among the 20 mission specialists. The candidates reported to the Johnson Space Center on July 1, 1978 to begin a challenging training and evaluation program that included Orbiter Systems training, science and enrichment briefings, and T-38 flight training. After successfully completing this program,

the following candidates were qualified as astronauts and received technical assignments within the Astronaut Office to prepare them further for an assignment to a Space Shuttle mission:

Guion S. Bluford, Jr., Maj., USAF. Born November 22, 1942. B.S. in Aerospace Engineering from Pennsylvania State University, M.S. in Aerospace Engineering from the Air Force Institute of Technology, and Ph.D. in Aerospace Engineering from the Air Force Institute of Technology.

Daniel C. Brandenstein, Lt. Comdr., USN. Born January 17, 1943. B.S. in Mathematics and Physics from the University of Wisconsin-River Falls.

James F. Buchli, Lt. Col., USMC. Born June 20, 1945. B.S. from U.S. Naval Academy and M.S. in Aeronautical Systems from the University of West Florida.

Michael L. Coats, Lt. Comdr., USN. Born January 16, 1946. B.S. from the U.S. Naval Academy, M.S. in the Administration of Science and Technology from George Washington University, and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Richard O. Covey, Maj., USAF. Born August 1, 1946. B.S. in Engineering Science from the U.S. Air Force Academy and M.S. in Aeronautics and Astronautics from Purdue University.

John O. Creighton, Lt. Comdr., USN. Born April 28, 1943. B.S. from the U.S. Naval Academy and M.S. in the Administration of Science and Technology from George Washington University.

John M. Fabian, Maj., USAF. Born January 28, 1939. B.S. in Mechanical Engineering from Washington State University, M.S. in Aerospace Engineering from the Air Force Institute of Technology, and Ph.D. in Aeronautics/Astronautics from the University of Washington.

Anna L. Fisher. Born August 24, 1949. B.S. in Chemistry from the University of California, Los Angeles and M.D. from the University of California, Los Angeles, School of Medicine.

Dale A. Gardner, Lt., USN. Born November 8, 1948. B.S. in Engineering Physics from the University of Illinois.

Robert L. Gibson, Lt., USN. Born October 30, 1946. B.S. in Aeronautical Engineering from California Polytechnic State University.

Frederick D. Gregory, Maj., USAF. Born January 7, 1941. B.S. from the U.S. Air Force Academy and M.S. in Information Systems from George Washington University.

Stanley D. Griggs. Born September 7, 1939. B.S. from U.S. Naval Academy and M.S.A. in Management Engineering from George Washington University.

Terry J. Hart. Born October 27, 1946. B.S. in Mechanical Engineering from Lehigh University, and M.S. in Mechanical Engineering from Massachusetts Institute of Technology.

Frederick H. Hauck, Comdr., USN. Born April 11, 1941. B.S. in General Physics from Tufts University and M.S. in Nuclear Engineering from Massachusetts Institute of Technology.

Steven A. Hawley. Born December 12, 1951. B.A. in Astronomy and Physics from the University of Kansas, and Ph.D. in Astronomy from the University of California, Santa Cruz.

Jeffrey A. Hoffman. Born November 2, 1944. B.A. in Astronomy from Amherst College, and Ph.D. in Astrophysics from Harvard University.

Shannon W. Lucid. Born January 14, 1943. B.S. in Chemistry from the University of Oklahoma, M.S. in Biochemistry from the University of Oklahoma, and Ph.D. in Biochemistry from the University of Oklahoma.

Jon A. McBride, Lt. Comdr., USN. Born August 14, 1943. B.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Ronald E. McNair. Born October 21, 1950. B.S. in Physics from the North Carolina A&T University and Ph.D. in Physics from Massachusetts Institute of Technology.

Richard M. Mullane, Capt., USAF. Born September 10, 1945. B.S. from the U.S. Military Academy and M.S. in Aeronautical Engineering from the Air Force Institute of Technology.

Steven R. Nagel, Capt., USAF. Born October 27, 1946. B.S. in Aeronautical/Astronautical Engineering from the University of Illinois.

George D. Nelson. Born July 13, 1950. B.S. in Physics from Harvey Mudd University, M.S. in Astronomy from the University of Washington, and Ph.D. in Astronomy from the University of Washington.

Ellison S. Onizuka, Capt., USAF. Born June 24, 1946. B.S. in Aerospace Engineering from the University of Colorado and M.S. in Aerospace Engineering from the University of Colorado.

Judith A. Resnik. Born April 5, 1949. B.S. in Electrical Engineering from Carnegie-Mellon University and Ph.D. in Electrical Engineering from the University of Maryland.

Sally K. Ride. Born May 26, 1951. B.S. in Physics from Stanford University, B.A. in English from Stanford University, and Ph.D. in Physics from Stanford University.

Francis R. Scobee, Maj., USAF. Born May 19, 1939. B.S. in Aerospace Engineering from the University of Arizona.

Margaret R. Seddon. Born November 8, 1947. B.A. in Physiology from the University of California, Berkeley, and M.D. from the University of Tennessee College of Medicine.

Brewster H. Shaw, Capt., USAF. Born May 16, 1945. B.S. in Engineering Mechanics from the University of Wisconsin and M.S. in Engineering Mechanics from the University of Wisconsin.

Loren J. Shriver, Capt., USAF. Born September 23, 1944. B.S. from the U.S. Air Force Academy and M.S. in Astronautics from Purdue University.

Robert L. Stewart, Maj., U.S. Army. Born August 13, 1942. B.S. in Mathematics from the University of Southern Mississippi and M.S. in Aerospace Engineering from the University of Texas, Arlington.

Kathryn D. Sullivan. Born October 3, 1951. B.S. in Earth Sciences from the University of California, Santa Cruz, and Ph.D. in Geology from Dalhousie University.

Norman E. Thagard. Born July 3, 1943. B.S. in Engineering Science from Florida State University, M.S. in Engineering Science from Florida State University, and M.D. from the University of Texas Southwestern Medical School.

James D. van Hoften. Born June 11, 1944. B.S. in Civil Engineering from the University of California, Berkeley, M.S. in Hydraulic Engineering from Colorado State University, and Ph.D. in Fluid Mechanics from Colorado State University.

David M. Walker, Lt. Comdr., USN. Born May 20, 1944. B.S. from the U.S. Naval Academy.

Donald E. Williams, Lt. Comdr., USN. Born February 13, 1942. B.S. in Mechanical Engineering from Purdue University.

The following individuals were on the rating and selection board for Group 8: George W.S. Abbey, Joseph D. Atkinson, Jr., Ph.D.; Vance D. Brand; Edward Gibson, Ph.D.; Carolyn Huntoon, Ph.D.; Joseph P. Kerwin, M.D.; Jack R. Lister; Glynn S. Lunney; Robert A. Parker; Robert O. Piland; Martin L. Raines; Duane L. Ross; Donald K. Slayton; James H. Trainor, Ph.D.; and John W. Young. Dr. James Tainor was from the NASA GSFC; the remainder were from NASA JSC.

Group 9

NASA selected 19 astronaut candidates for the Space Shuttle program in 1980. The eight pilots and 11 mission specialists included the first Hispanic mission specialist and the first candidate selected from the U.S. Army. The candidates reported to the Johnson Space Center on July 7, 1980 to begin their training and evaluation program. The 1980 astronaut candidate class included:

James P. Bagian. Born February 22, 1952. B.S. in Mechanical Engineering from Drexel University and M.D. from Thomas Jefferson University.

John E. Blaha, Lt. Col., USAF. Born August 26, 1942. B.S. in Astronautical Engineering from the U.S. Air Force Academy and M.S. in Astronautical Engineering from Purdue University.

Charles F. Bolden, Jr., Maj., USMC. Born August 19, 1946. B.S. in Electrical Engineering from the U.S. Naval Academy and M.S. in Systems Management from the University of California.

Roy D. Bridges, Jr., Maj., USAF. Born July 19, 1943. B.S. in Engineering Science from the U.S. Air Force Academy and M.S. in Astronautics from Purdue University.

Franklin R. Chang. Born April 5, 1950. B.S. in Mechanical Engineering from the University of Connecticut and Ph.D. in Physics from Massachusetts Institute of Technology.

- Mary L. Cleave. Born February 5, 1947. B.S. in Biology from Colorado State University, M.S. in Botany from Utah State University, and Ph.D. in Civil Engineering from Utah State University.
- Bonnie J. Dunbar. Born March 3, 1949. B.S. in Ceramic Engineering from the University of Washington and M.S. in Ceramic Engineering from the University of Washington.
- William F. Fisher. Born April 1, 1946. B.S. in Psychology from Stanford University, M.S. in Engineering Science from the University of Houston and M.D. from the University of Florida College of Medicine.
- Guy S. Gardner, Maj., USAF. Born January 6, 1948. B.S. in Aeronautical Engineering from the U.S. Air Force Academy and M.S. in Aeronautical Engineering from Purdue University.
- Ronald J. Grabe, Maj., USAF. Born June 13, 1945. B.S. in Engineering Science from the U.S. Air Force Academy and M.S. in Aeronautics from the Technische Hochschule, Darmstadt, Germany.
- David C. Hilmers, Capt., USMC. Born January 28, 1950. B.S. in Mathematics from Cornell College and M.S. in Electrical Engineering from the U.S. Naval Postgraduate School.
- David C. Leestma, Lt. Comdr., USN. Born May 6, 1949. B.S. in Aeronautical Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.
- John M. Lounge. Born June 28, 1946. B.S. in Mathematics from the U.S. Naval Academy and M.S. in Astrophysics from the University of Colorado.
- Bryan D. O'Connor, Maj., USMC. Born September 6, 1946. B.S. in Naval Science from the U.S. Naval Academy and M.S. in Aeronautical Systems from the University of West Florida.
- Richard N. Richards, Lt. Comdr., USN. Born August 24, 1946. B.S. in Chemical Engineering from the University of Missouri and M.S. in Aeronautical Systems from the University of West Florida.
- Jerry L. Ross, Capt., USAF. Born January 20, 1948. B.S. in Mechanical Engineering from Purdue University and M.S. in Mechanical Engineering from Purdue University.
- Michael J. Smith, Lt. Comdr., USN. Born April 30, 1945. B.S. in Aeronautical Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.
- Sherwood C. Spring, Maj., U.S. Army. Born September 3, 1944. B.S. in Engineering from the U.S. Military Academy and M.S. in Aerospace Engineering from the University of Arizona.
- Robert C. Springer, Maj., USMC. Born May 21, 1942. B.S. in Naval Science from the U.S. Naval Academy, and M.S. in Operations Research from the U.S. Naval Postgraduate School.

The following individuals served on the rating panel and the astronaut candidate selection board for Group 9: George W.S. Abbey; Joseph D. Atkinson, Jr., Ph.D.; Joseph P. Allen, IV, Ph.D.; Vance D. Brand; Harvey L. Hartman; Gregory W.

Hayes; Jay F. Honeycutt; Carolyn L. Huntoon Ph.D.; William B. Lenoir, Ph.D.; Robert O. Piland; James H. Trainor, Ph.D.; Paul J. Weitz; and John W. Young. Dr. Trainor was from NASA GSFC, and the remainder were from NASA JSC.

Group 10

In 1984, NASA selected 17 astronaut candidates for the Space Shuttle program. One of the seven pilots was Hispanic, and three of the 10 mission specialists were female. The astronaut candidates who began training at the Johnson Space Center on July 1, 1984 included:

James C. Adamson, Maj., U.S. Army. Born March 3, 1946. B.S. in Engineering from the U.S. Military Academy and M.S. in Aeronautical and Mechanical Engineering from Princeton University.

Mark N. Brown, Capt., USAF. Born November 18, 1951. B.S. in Aeronautical and Astronautical Engineering from Purdue University and M.S. in Engineering from the Air Force Institute of Technology.

Kenneth D. Cameron, Maj., USMC. Born November 29, 1949. B.S. in Aeronautics and Astronautics from Massachusetts Institute of Technology and M.S. in Aeronautics and Astronautics from Massachusetts Institute of Technology.

Manley L. Carter, Jr., Comdr., USN. Born August 15, 1947. B.A. in Chemistry from Emory University and M.D. from Emory University.

John H. Casper, Lt. Col., USAF. Born July 9, 1943. B.S. in Astronautics and Engineering Science from the U.S. Airforce Academy and M.S. in Astronautics from Purdue University.

Frank L. Culbertson, Jr., Lt. Comdr., USN. Born May 15, 1949. B.S. in Aerospace Engineering from the U.S. Naval Academy.

Sidney M. Gutierrez, Capt., USAF. Born June 27, 1951. B.S. in Aerospace Engineering from the U.S. Air Force Academy and M.A. in Management from Webster College.

Lloyd B. Hammond, Jr., Capt., USAF. Born January 16, 1952. B.S. in Engineering Mechanics from the U.S. Air Force Academy and M.S. in Engineering Mechanics from Georgia Institute of Technology.

Marsha S. Ivins. Born April 15, 1951. B.S. in Aerospace Engineering from the University of Colorado.

Mark C. Lee, Capt., USAF. Born August 14, 1952. B.S. in Civil Engineering from the U.S. Air Force Academy and M.S. in Mechanical Engineering from Massachusetts Institute of Technology.

George D. Low. Born February 19, 1956. B.S. in Physics from Washington and Lee University, B.S. in Mechanical Engineering from Cornell University and M.S. in Aeronautics and Astronautics from Stanford University.

Michael J. McCulley, Lt. Comdr., USN. Born August 4, 1943. B.S. in Metallurgical Engineering from Purdue University and M.S. in Metallurgical Engineering from Purdue University.

William M. Shepherd, Lt. Comdr., USN. Born July 26, 1949. B.S. in Aerospace Engineering from the U.S. Naval Academy, M.S. in Mechanical Engineering from Massachusetts Institute of Technology, and M.S. in Ocean Engineering from Massachusetts Institute of Technology.

Ellen L. Shulman. Born April 27, 1953. B.A. in Geology from the State University of New York at Buffalo and M.D. from Cornell University.

Kathryn C. Thornton. Born August 17, 1952. B.S. in Physics from Auburn University, M.S. in Physics from the University of Virginia, and Ph.D. in Physics from the University of Virginia.

Charles L. Veach. Born September, 18, 1944. B.S. in Engineering Management from the U.S. Air Force Academy.

James D. Wetherbee, Lt., USN. Born November 27, 1952. B.S. in Aerospace Engineering from the University of Notre Dame.

The members of the rating panel and selection committee for Group 10 included George W.S. Abbey; Joseph D. Atkinson, Jr., Ph.D.; Joseph P. Allen IV, Ph.D.; Gregory W. Hayes; Jay F. Honeycutt; Carolyn L. Huntoon, Ph.D.; Robert A.R. Parker, Ph.D.; Duane L. Ross; Paul J. Weitz; and John W. Young. All were from NASA JSC.

Group 11

An additional 13 candidates were selected to train for the astronaut corps in 1985. This group of six pilots and seven mission specialists included two females. They reported to the Johnson Space Center on August 1, 1985. The astronaut candidates were:

Jerome Apt. Born April 28, 1949. B.S. in Physics from Harvard College and Ph.D. in Physics from the Massachusetts Institute of Technology.

Michael A. Baker, Lt. Comdr., USN. Born October 27, 1953. B.S. in Aerospace Engineering from the University of Texas.

Robert D. Cabana, Maj., USMC. Born January 23, 1949. B.S. in Mathematics from the U.S. Naval Academy.

Brian Duffy, Capt., USAF. Born June 20, 1953. B.S. in Mathematics from the U.S. Air Force Academy and M.S. in Systems Management from the University of Southern California.

Charles D. Gemar, Capt., U.S. Army. Born August 4, 1955. B.S. in Engineering from the U.S. Military Academy.

Linda M. Godwin. Born July 2, 1952. B.S. in Mathematics from Southeast Missouri State, M.S. in Physics from the University of Missouri, and Ph.D. in Physics from the University of Missouri.

Terence T. Henricks, Maj., USAF. Born July 5, 1952. B.S. in Civil Engineering from the U.S. Air Force Academy and M.S. in Public Administration from Golden Gate University.

Richard J. Hieb. Born Sept 21, 1955. B.S. in Mathematics and Physics from Northwest Nazarene College and M.S. in Aerospace Engineering from the University of Colorado.

Tamara E. Jernigan. Born May 7, 1959. B.S. in Physics from Stanford University, M.S. in Engineering Science from Stanford University, and M.S. in Astronomy from the University of California at Berkeley.

Carl J. Meade, Capt., USAF. Born November 16, 1950. B.S. in Electronics Engineering from the University of Texas and M.S. in Electronics Engineering from California Institute of Technology.

Stephen S. Oswald. Born June 30, 1951. B.S. in Aerospace Engineering from the U.S. Naval Academy.

Stephen D. Thorne, Lt. Comdr., USN. Born February 11, 1953. B.S. in Engineering from the U.S. Naval Academy.

Pierre J. Thuot, Lt., USN. Born May 19, 1955. B.S. in Physics from the U.S. Naval Academy and M.S. in Systems Management from the University of Southern California.

The Group 11 rating panel and selection committee members were George W.S. Abbey; Joseph D. Atkinson, Jr., Ph.D.; Karol J. Bobko; Daniel C. Brandenstein; Mary L. Cleave, Ph.D.; Richard O. Covey; Anna L. Fisher, M.D.; David C. Leestma; George D. Nelson, Ph.D.; Ellison S. Onizuka; Sally K. Ride, Ph.D.; Duane L. Ross; Loren J. Shriver; Paul J. Weitz; Donald E. Williams; and John W. Young. All were from NASA JSC.

Group 12

The 15 astronaut candidates selected in 1987 consisted of seven pilots and eight mission specialists, including the first African-American female and the first candidate from the U.S. Coast Guard. The following astronaut candidates reported to Johnson Space Center on August 17, 1987, to begin training:

Thomas D. Akers, Capt., USAF. Born May 20, 1951. B.S. in Applied Mathematics from the University of Missouri-Rolla and M.S. in Applied Mathematics from the University of Missouri-Rolla.

Andrew M. Allen, Capt., USMC. Born August 4, 1955. B.S. in Mechanical Engineering from Villanova University.

Kenneth D. Bowersox, Lt., USN. Born November 14, 1956. B.S. in Aerospace Engineering from the U.S. Naval Academy and M.S. in Mechanical Engineering from Columbia University.

Curtis L. Brown, Capt., USAF. Born March 11, 1956. B.S. in Electrical Engineering from the U.S. Air Force Academy.

Kevin P. Chilton, Maj., USAF. Born November 3, 1954. B.S. in Engineering Science from the U.S. Air Force Academy and M.S. in Engineering Mechanics from Columbia University.

Jan D. Dozier. Born November 1, 1953. B.S. in Biology from the Georgia Institute of Technology, B.S. in Mechanical Engineering from Auburn University, M.S. in Mechanical Engineering from the University of Alabama and Ph.D. in Mechanical Engineering from the University of Alabama.

C. Michael Foale. Born January 6, 1957. B.A. in Physics from Cambridge University, M.A. in Physics from Cambridge University and Ph.D. in Physics from Cambridge University (England).

Gregory J. Harbaugh. Born January 15, 1956. B.S. in Aeronautical and Astronautical Engineering from Purdue University and M.S. in Physical Science from the University of Houston-Clear Lake.

Mae C. Jemison. Born October 17, 1956. B.S. in Chemical Engineering from Stanford University and M.D. from Cornell University.

Donald R. McMonagle, Maj., USAF. Born May 14, 1952. B.S. in Astronautical Engineering from the U.S. Air Force Academy and M.S. in Mechanical Engineering from California State University-Fresno.

Bruce E. Melnick, Lt. Comdr., USCG. Born December 5, 1949. B.S. in Ocean Engineering from the U.S. Coast Guard Academy and M.S. in Aeronautical Systems from the University of West Florida.

William F. Readdy, Lt. Comdr., USN. Born January 24, 1952. B.S. in Aeronautical Engineering from the U.S. Naval Academy.

Kenneth S. Reightler, Jr., Lt. Comdr., USN. Born March 24, 1951. B.S. in Aerospace Engineering from the U.S. Naval Academy, M.S. in Systems Management from the University of Southern California, and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Mario Runco, Jr., Lt. Comdr., USN. Born January 26, 1952. B.S. in Meteorology from City College of New York and M.S. in Meteorology from Rutgers University.

James S. Voss, Maj., U.S. Army. Born March 3, 1949. B.S. in Aerospace Engineering from Auburn University and M.S. in Aerospace Engineering from the University of Colorado.

The following individuals served on the rating panel and selection board for Group 12: George W.S. Abbey; Joseph D. Atkinson, Jr., Ph.D.; Daniel C. Brandenstein; Anna L. Fisher, M.D.; Carolyn L. Huntoon, Ph.D.; David C. Leestma; George D. Nelson, Ph.D.; Duane L. Ross; Loren J. Shriver; Paul J. Weitz; and John W. Young. All were from NASA JSC.

Group 13

The 23 astronaut candidates selected in 1990 included 7 pilots and 16 mission specialists. Among the candidates was the first female selected as a pilot and the first Hispanic female selected as a mission specialist. The following astronaut candidates began training at the Johnson Space Center on July 15, 1990:

Daniel W. Bursch, Lt. Comdr., USN. Born July 25, 1957. B.S. in Physics from the U.S. Naval Academy.

- Leroy Chiao. Born August 28, 1960. B.S. in Chemical Engineering from the University of California-Berkeley, M.S. in Chemical Engineering from the University of California-Santa Barbara, and Ph.D. in Chemical Engineering from the University of California-Santa Barbara.
- Michael R.U. Clifford, Maj., U.S. Army. Born October 13, 1952. B.S. in Basic Science from the U.S. Military Academy and M.S. in Aerospace Engineering from the Georgia Institute of Technology.
- Kenneth D. Cockrell. Born April 9, 1950. B.S. in Mechanical Engineering from the University of Texas and M.S. in Aeronautical Systems from the University of Florida.
- Eileen M. Collins, Maj., USAF. Born November 19, 1956. B.A. in Mathematics from Syracuse University, M.S. in Operations Research from Stanford University, and M.A. in Space Systems Management from Webster University.
- William G. Gregory, Capt., USAF. Born May 14, 1957. B.S. in Engineering Science from the U.S. Air Force Academy, M.S. in Engineering Mechanics from Columbia University, and M.S. in Management from Troy State.
- James D. Halsell, Maj., USAF. Born September 29, 1956. B.S. in Engineering from the U.S. Air Force Academy, M.S. in Management from Troy State, and M.S. in Space Operations from the Air Force Institute of Technology.
- Bernard A. Harris, Jr. Born June 26, 1956. B.S. in Biology from the University of Houston and M.D. from Texas Tech University.
- Susan J. Helms, Capt., USAF. Born February 26, 1958. B.S. in Aerospace Engineering from the U.S. Air Force Academy and M.S. in Aeronautics/Astronautics from Stanford University.
- Thomas D. Jones. Born January 22, 1955. B.S. in Basic Science from the U.S. Air Force Academy and Ph.D. in Planetary Science from the University of Arizona.
- William S. McArthur, Jr., Maj., U.S. Army. Born July 26, 1951. B.S. in Applied Sciences and Engineering from the U.S. Military Academy and M.S. in Aerospace Engineering from the Georgia Institute of Technology.
- James H. Newman. Born October 16, 1956. B.A. in Physics from Dartmouth College, M.A. in Physics from Rice University, and Ph.D. in Physics from Rice University.
- Ellen Ochoa. Born May 10, 1958. B.S. in Physics from San Diego State, M.S. in Electrical Engineering from Stanford University, and Ph.D. in Electrical Engineering from Stanford University.
- Charles J. Precourt, Maj., USAF. Born June 29, 1955. B.S. in Aeronautical Engineering from the U.S. Air Force Academy and M.S. in Management from Golden Gate University.
- Richard A. Searfoss, Maj., USAF. Born June 6, 1956. B.S. in Aerospace Engineering from the U.S. Air Force Academy and M.S. in Aerospace Engineering from the California Institute of Technology.
- Ronald M. Sega. Born December 4, 1952. B.S. in Physics and Mathematics from the U.S. Air Force Academy, M.S. in Physics from Ohio State, and Ph.D. in Electrical Engineering from the University of Colorado.

Nancy J. Sherlock, Capt., U.S. Army. Born December 29, 1958. B.A. in Biological Science from Ohio State and M.S. in Safety Engineering from the University of Southern California.

Donald A. Thomas. Born May 6, 1955. B.S. in Physics from Case Western University, M.S. in Materials Science from Cornell University, and Ph.D. in Materials Science from Cornell University.

Janice E. Voss. Born October 8, 1956. B.S. in Engineering Science from Purdue University, M.S. in Electrical Engineering from Massachusetts Institute of Technology, and Ph.D. in Aeronautics/Astronautics from Massachusetts Institute of Technology.

Carl E. Walz, Capt., USAF. Born September 6, 1955. B.S. in Physics from Kent State and M.S. in Physics from John Carroll University.

Terrence W. Wilcutt, Maj., USMC. Born October 31, 1949. B.A. in Mathematics from Western Kentucky University.

Peter J. K. Wisoff. Born August 16, 1958. B.S. in Physics from the University of Virginia, M.S. in Physics from Stanford University, and Ph.D. in Applied Physics from Stanford University.

David A. Wolf. Born August 23, 1956. B.S. in Electrical Engineering from Purdue University and M.D. from Indiana University.

The rating panel and selection board members for the 1990 astronaut candidate class were Joseph D. Atkinson, Jr., Ph.D.; Charles F. Bolden; Daniel C. Brandenstein; Mary L. Cleave, Ph.D.; Michael L. Coats; Richard O. Covey; Steven A. Hawley, Ph.D.; Jeffrey A. Hoffman, Ph.D.; Carolyn L. Huntoon, Ph.D.; Robert A.R. Parker, Ph.D.; Donald R. Puddy; Duane L. Ross; Jerry L. Ross; Rhea Seddon, M.D.; and John W. Young. All were from NASA JSC.

Group 14

In 1992, NASA selected 19 new astronaut candidates in support of the Space Shuttle program. The group consisted of four pilots and 15 mission specialists, including nine civilians and 10 military officers. The following astronaut candidates reported to the Johnson Space Center on August 3, 1992 to begin their training and evaluation program:

Daniel T. Barry. Born December 30, 1953. B.S. in Electrical Engineering from Cornell University, M.S.E. in Electrical Engineering and Computer Science from Princeton University, M.A. in Electrical Engineering and Computer Science from Princeton University, Ph.D. in Electrical Engineering and Computer Science from Princeton University, and M.D. from the University of Miami.

Charles E. Brady, Jr., M.D., Comdr., USN. Born August 12, 1951. Premed at the University of North Carolina at Chapel Hill and M.D. from Duke University.

- Catherine G. Coleman, Ph.D., Capt., USAF. Born December 14, 1960. B.S. in Chemistry from Massachusetts Institute of Technology and Ph.D. in Polymer Science and Engineering from the University of Massachusetts.
- Michael L. Gernhardt. Born May 4, 1956. B.S. in Physics from Vanderbilt University, M.S. in Bioengineering from the University of Pennsylvania, and Ph.D. in Bioengineering from the University of Pennsylvania.
- John M. Grunsfeld. Born October 10, 1958. B.S. in Physics from the Massachusetts Institute of Technology, M.S. in Physics from the University of Chicago, and Ph.D. in Physics from the University of Chicago.
- Scott J. Horowitz, Ph.D., Capt., USAF. Born March 24, 1957. B.S. in Engineering from California State University at Northridge, M.S. in Aerospace Engineering from Georgia Tech, and Ph.D. in Aerospace Engineering from the Georgia Institute of Technology.
- Brent W. Jett, Jr., Lt. Comdr., USN. Born October 5, 1958. B.S. in Aerospace Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.
- Kevin R. Kregel. Born September 16, 1956. B.S. in Astronautical Engineering from the U.S. Air Force Academy and M.P.A. in Public Administration from Troy State University.
- Wendy B. Lawrence, Lt. Comdr., USN. Born July 2, 1959. B.S. in Ocean Engineering from the U.S. Naval Academy and M.S. in Ocean Engineering from Massachusetts Institute of Technology.
- Jerry M. Linenger, Comdr., USN. Born January 16, 1955. B.S. in Bioscience from the U.S. Naval Academy, M.D. from Wayne State University, M.S. in Systems Management, M.P.H. from the University of North Carolina, and Ph.D. in Epidemiology from the University of North Carolina.
- Richard M. Linnehan, D.V.M., Capt., U.S. Army. Born September 19, 1957. B.S. in Zoology from the University of New Hampshire and D.V.M. from Ohio State University.
- Michael E. Lopez-Alegria, Lt. Comdr., USN. Born May 30, 1958. B.S. in Systems Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.
- Scott E. Parazynski. Born July 28, 1961. B.S. in Biology from Stanford University and M.D. from Stanford University.
- Kent V. Rominger, Lt. Comdr., USN. Born August 7, 1956. B.S. in Civil Engineering from Colorado State University and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.
- Winston E. Scott, Comdr., USN. Born August 6, 1950. B.A. in Music from Florida State University and M.S. in Aeronautical Engineering from the U.S. Naval Academy.
- Steven L. Smith. Born December 30, 1958. B.S. in Electrical Engineering from Stanford University, M.S. in Electrical Engineering from Stanford University, and M.B.A. from Stanford University.
- Joseph R. Tanner. Born January 21, 1950. B.S. in Mechanical Engineering from the University of Illinois.

Andrew S.W. Thomas. Born December 18, 1951. B.E. in Mechanical Engineering from the University of Adelaide (Australia) and Ph.D. in Mechanical Engineering from the University of Adelaide (Australia).

Mary E. Weber. Born August 24, 1962. B.S. in Chemical Engineering from Purdue University and Ph.D. in Chemistry from the University of California-Berkeley.

The following persons served on the rating panel and astronaut candidate selection board in 1991: Thomas D. Akers; Joseph D. Atkinson, Jr., Ph.D.; Ellen S. Baker, M.D.; Robert D. Cabana; Franklin R. Chang-Diaz, Ph.D.; Richard O. Covey; Bonnie J. Dunbar, Ph.D.; Robert L. Gibson; Linda M. Godwin, Ph.D.; Jeffrey A. Hoffman, Ph.D.; Carolyn L. Huntoon, Ph.D.; Roger L. Kroes, Ph.D.; David C. Leestma; Paul Lowman, Ph.D.; Donald R. McMonagle; Donald R. Puddy; Duane L. Ross; Rhea Seddon, M.D.; William M. Shepherd; Loren J. Shriver; Kathryn D. Sullivan, Ph.D.; Kathryn C. Thornton, Ph.D.; James D. Wetherbee; and John W. Young. Dr. Kroes was from NASA Marshall Space Flight Center (MSFC) and Dr. Lowman was from NASA GSFC. The others were from NASA JSC.

Group 15

The 1995 astronaut candidate class consisted of 10 pilot candidates, including two female pilots, and nine mission specialists. The 19 astronaut candidates were selected through a highly competitive process that evaluated their education, experience, and ability to work as members of a team. This was the first astronaut class to receive training on both the Space Shuttle and the International Space Station programs. The 1995 astronaut candidates included:

Scott D. Altman, Lt. Comdr., USN. Born August 15, 1959. B.S. in Aeronautical and Astronautical Engineering from the University of Illinois and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Michael P. Anderson, Maj., USAF. Born December 25, 1959. B.S. in Physics and Astronomy from the University of Washington and M.S. in Physics from Creighton University.

Jeffrey S. Ashby, Comdr., USN. Born June 16, 1954. B.S. in Mechanical Engineering from the University of Idaho and M.S. in Aviation Systems from the University of Tennessee.

Michael J. Bloomfield, Maj., USAF. Born March 16, 1959. B.S. in Engineering Mechanics from the U.S. Air Force Academy and M.S. in Engineering Management from Old Dominion University.

Kalpana Chawla. Born July 1, 1961. B.S. in Aeronautical Engineering from Punjab Engineering College in India, M.S. in Aerospace Engineering from the University of Texas, and Ph.D. in Aerospace Engineering from the University of Colorado.

Robert Curbeam, Jr., Lt. Comdr., USN. Born March 5, 1962. B.S. in Aerospace Engineering from the U.S. Naval Academy, M.S. in Aeronautical Engineering

from the U.S. Naval Postgraduate School, and Degree of Aeronautical and Astronautical Engineer from the U.S. Naval Postgraduate School.

Joe F. Edwards, Jr., Lt. Comdr., USN. Born February 3, 1958. B.S. in Aerospace Engineering from the U.S. Naval Academy and M.S. in Aviation from the University of Tennessee at Knoxville.

Dominic L. Gorie, Comdr., USN. Born May 2, 1957. B.S. in Ocean Engineering from the U.S. Naval Academy and M.S. in Aviation Systems from the University of Tennessee at Knoxville.

Kathryn P. Hire. Born August 26, 1959. B.S. in Engineering Management from the U.S. Naval Academy and M.S. in Space Technology from the Florida Institute of Technology.

Rick D. Husband, Maj., USAF. Born July 12, 1957. B.S. in Mechanical Engineering from Texas Tech University and M.S. in Mechanical Engineering from California State University.

Janet L. Kavandi. Born July 17, 1959. B.S. in Chemistry from Missouri Southern State College, M.S. in Chemistry from the University of Missouri-Rolla, and Ph.D. in Analytical Chemistry from the University of Washington.

Steven W. Lindsey, Maj., USAF. Born August 24, 1960. B.S. in Engineering Sciences from the U.S. Air Force Academy and M.S. in Aerospace Engineering from the Air Force Institute of Technology.

Edward T. Lu. Born July 1, 1963. B.S. in Electrical Engineering from Cornell University and Ph.D. in Applied Physics from Stanford University.

Pamela A. Melroy, Maj., USAF. Born September 17, 1961. B.S. in Physics and Astronomy from Wellesley College and M.S. in Earth and Planetary Sciences from the Massachusetts Institute of Technology.

Carlos I. Noriega, Maj., USMC. Born October 8, 1959. B.S. in Computer Science from the University of Southern California, M.S. in Computer Science from the U.S. Naval Postgraduate School, and M.S. in Space Systems Operations from the U.S. Naval Postgraduate School.

James F. Reilly. Born March 18, 1954. B.S. in Geosciences from the University of Texas, M.S. in Geosciences from the University of Texas, and Ph.D. in Geosciences from the University of Texas.

Stephen K. Robinson. Born October 26, 1955. B.S. in Mechanical and Aeronautical Engineering from the University of California, M.S. in Mechanical Engineering from Stanford University and Ph.D. in Mechanical Engineering from Stanford University.

Susan L. Still, Lt., USN. Born October 24, 1961. B.S. in Aeronautical Engineering from Embry-Riddle University and M.S. in Aerospace Engineering from Georgia Institute of Technology.

Frederick W. Sturckow, Capt., USMC. Born August 11, 1961. B.S. in Mechanical Engineering from California Polytechnic State University.

The 1994 rating panel and astronaut candidate selection board members were Thomas D. Akers; Joseph D. Atkinson, Jr., Ph.D.; Ellen S. Baker, M.D.; Kenneth D. Cameron; Kevin P. Chilton; Michael R. Clifford; Brian Duffy;

Michael Foale, Ph.D.; Robert L. Gibson; Estella H. Gillette; Linda M. Godwin, Ph.D.; Frederick D. Gregory; James F. Harrington; Bernard A. Harris, M.D.; Steven A. Hawley, Ph.D.; Gregory W. Hayes; Carolyn L. Huntoon, Ph.D.; Tamara E. Jernigan, Ph.D.; Roger L. Kroes, Ph.D.; Mark C. Lee; David C. Leestma; Paul D. Lowman, Ph.D.; David H. Mobley; John F. Muratore; Steven R. Nagel; Ellen Ochoa, Ph.D.; Stephen S. Oswald; Duane L. Ross; Jerry L. Ross; Pierre J. Thuot; James S. Voss; James D. Wetherbee; and John W. Young. Mr. Harrington was from the NASA KSC, Dr. Kroes was from NASA MSFC, and Dr. Lowman was from NASA GSFC. The others were from NASA JSC.

Group 16

In 1996, NASA selected 35 astronaut candidates in support of the Space Shuttle and Space Station programs. This was the largest astronaut candidate class since the first group of Shuttle astronauts was selected in 1978. The 10 pilots and 25 mission specialists included three African-American females and a pair of identical twins. On 12 August 1996, the candidates began a curriculum including extensive Space Shuttle and Space Station training. Upon successful completion of their training, they were qualified as astronauts and began supporting long-duration missions on the International Space Station. The 1996 candidates included:

David M. Brown, Comdr., USN. Born April 16, 1956. B.S. in Biology from the College of William and Mary and M.D. from Eastern Virginia Medical School.

Daniel C. Burbank, Lt. Comdr., USCG. Born July 27, 1961. B.S. in Electrical Engineering from the U.S. Coast Guard Academy and M.S. in Aeronautical Science from Embry-Riddle Aeronautical University.

Yvonne D. Cagle. Born April 24, 1959. B.A. in Biochemistry from San Francisco State University and M.D. from the University of Washington.

Fernando Caldeiro. Born June 12, 1958. B.S. in Mechanical Engineering from the University of Arizona and M.S. in Engineering Management from the University of Central Florida.

Charles J. Camarda. Born May 8, 1952. B.S. in Aerospace Engineering from Polytechnic Institute of New York, M.S. in Engineering Science from George Washington University, and Ph.D. in Aerospace Engineering from Virginia Polytechnic Institute.

Duane G. Carey, Maj., USAF. Born April 30, 1957. B.S. in Aerospace Engineering and Mechanics from the University of Minnesota-Minneapolis and M.S. in Aerospace Engineering from the University of Minnesota-Minneapolis.

Laurel B. Clark, Lt. Comdr., USN. Born March 10, 1961. B.S. in Zoology from the University of Wisconsin-Madison and M.D. from the University of Wisconsin-Madison.

Edward M. Fincke, Capt., USAF. Born March 14, 1967. B.S. in Aeronautics and Astronautics, and Earth, Atmospheric, and Planetary Sciences from

Massachusetts Institute of Technology, and M.S. in Aeronautics and Astronautics from Stanford University.

Patrick G. Forrester, Lt. Col., U.S. Army. Born March 31, 1957. B.S. in Applied Sciences and Engineering from the U.S. Military Academy and M.S. in Mechanical and Aerospace Engineering from the University of Virginia.

Stephen N. Frick, Lt. Comdr., USN. Born September 30, 1964. B.S. in Aerospace Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

John B. Herrington, Lt. Comdr., USN. Born September 14, 1958. B.S. in Applied Mathematics from the University of Colorado and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Joan E. Higginbotham. Born August 3, 1964. B.S. in Electrical Engineering from Southern Illinois, M.S. in Management from Florida Institute of Technology, and M.S. in Space Systems from Florida Institute of Technology.

Charles O. Hobaugh, Capt., USMC. Born November 5, 1961. B.S. in Aerospace Engineering from the U.S. Naval Academy.

James M. Kelly, Capt., USAF. Born May 14, 1964. B.S. in Astronautical Engineering from the U.S. Air Force Academy.

Mark E. Kelly, Lt., USN. Born February 21, 1964. B.S. in Marine Engineering and Nautical Science from the U.S. Merchant Marine Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Scott J. Kelly, Lt., USN. Born February 21, 1964. B.S. in Electrical Engineering from the State University of New York Maritime College.

Paul S. Lockhart, Maj., USAF. Born April 28, 1956. B.A. in Mathematics from Texas Tech University and M.S. in Aerospace Engineering from the University of Texas.

Christopher J. Loria, Maj., USMC. Born July 9, 1960. B.S. in General Engineering from the U.S. Naval Academy.

Sandra H. Magnus. Born October 30, 1964. B.S. in Physics from the University of Missouri-Rolla, M.S. in Electrical Engineering from the University of Missouri-Rolla, and Ph.D. in Materials Science and Engineering from Georgia Institute of Technology.

Michael J. Massimino. Born August 19, 1962. B.S. in Industrial Engineering from Columbia University, M.S. in Mechanical Engineering and Technology and Policy from Massachusetts Institute of Technology, and Ph.D. in Mechanical Engineering from Massachusetts Institute of Technology.

Richard A. Mastracchio. Born February 11, 1960. B.S. in Electrical Engineering and Computer Science from the University of Connecticut, M.S. in Electrical Engineering from Rensselaer Polytechnic Institute, and M.S. in Physical Sciences from the University of Houston-Clear Lake.

William C. McCool, Lt. Comdr., USN. Born September 23, 1961. B.S. in Applied Science from the U.S. Naval Academy, M.S. in Computer Science from the University of Maryland, and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Lee M. Morin, Comdr., USN. Born September 9, 1952. B.S. in Mathematics and Electrical Science from the University of New Hampshire, M.S. in Biochemistry from New York University, M.D. from New York University, Ph.D. in Microbiology from New York University, and M.P.H. from the University of Alabama at Birmingham.

Lisa M. Nowak, Lt. Comdr., USN. Born May 10, 1963. B.S. in Aerospace Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Donald R. Pettit. Born April 20, 1955. B.S. in Chemical Engineering from Oregon State University and Ph.D. in Chemical Engineering from the University of Arizona.

John L. Phillips. Born April 15, 1951. B.S. in Mathematics and Russian from the U.S. Naval Academy, M.S. in Aeronautical Systems from the University of West Florida, M.S. in Geophysics and Space Physics from the University of California, and Ph.D. in Geophysics and Space Physics from the University of California.

Mark L. Polansky. Born June 2, 1956. B.S. in Aeronautical and Astronautical Engineering from Purdue University and M.S. in Aeronautics and Astronautics from Purdue University.

Paul W. Richards. Born May 20, 1964. B.S. in Mechanical Engineering from Drexel University and M.S. in Mechanical Engineering from the University of Maryland.

Piers J. Sellers. Born April 11, 1955. B.S. in Ecological Science from the University of Edinburgh (Scotland) and Ph.D. in Biometeorology from Leeds University (United Kingdom).

Heidemarie M. Stefanyshyn-Piper, Lt. Comdr., USN. Born February 7, 1963. B.S. in Mechanical Engineering from Massachusetts Institute of Technology and M.S. in Mechanical Engineering from Massachusetts Institute of Technology.

Daniel M. Tani. Born February 1, 1961. B.S. in Mechanical Engineering from Massachusetts Institute of Technology and M.S. in Mechanical Engineering from Massachusetts Institute of Technology.

Rex J. Walheim, Capt., USAF. Born October 10, 1962. B.S. in Mechanical Engineering from the University of California-Berkeley and M.S. in Industrial Engineering from the University of Houston.

Peggy A. Whitson. Born February 9, 1960. B.S. in Biology and Chemistry from Iowa Wesleyan College and Ph.D. in Biochemistry from Rice University.

Jeffrey N. Williams, Maj., U.S. Army. Born January 18, 1958. B.S. in Applied Sciences and Engineering from U.S. Military Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Stephanie D. Wilson. Born September 27, 1966. B.S. in Engineering Science from Harvard University and M.S. in Aerospace Engineering from the University of Texas.

The 1995 astronaut candidate rating panel and selection board included the following individuals: Thomas D. Akers; Joseph D. Atkinson, Jr., Ph.D.;

Ellen S. Baker, M.D.; Robert D. Cabana; Brian Duffy; Michael Foale, Ph.D.; Estella H. Gillette; Linda M. Godwin, Ph.D.; Steven A. Hawley, Ph.D.; Gregory W. Hayes; Susan J. Helms; Robert K. Holkan; Donald R. McMonagle; David H. Mobley; John F. Muratore; Bascom W. Murrah III; Duane L. Ross; John A. Rummel, Ph.D.; James D. Wetherbee; and John W. Young. Mr. Mobley was from NASA Headquarters, and Mr. Murrah was from NASA KSC. The others were from NASA JSC.

Group 17

In 1998, NASA selected 25 astronaut candidates to support the Space Station and Space Shuttle programs. The group of 8 pilots and 17 mission specialists consisted of 21 males and 4 females, including NASA's first Educator Mission Specialist. The astronaut class reported for training at the Johnson Space Center on August 15, 1998. The candidates were:

Clayton C. Anderson. Born February 23, 1959. B.S. in Physics from Hastings College and M.S. in Aerospace Engineering from Iowa State University.

Lee J. Archambault, Maj., USAF. Born August 25, 1960. B.S. in Aerospace and Astronautical Engineering from the University of Illinois-Urbana and M.S. in Aerospace and Astronautical Engineering from the University of Illinois-Urbana.

Tracy E. Caldwell. Born August 14, 1969. B.S. in Chemistry from California State University-Fullerton and Ph.D. in Chemistry from the University of California-Davis.

Gregory E. Chamitoff. Born August 6, 1962. B.S. in Electrical Engineering from California Polytechnic State University, M.S. in Aerospace Engineering from California Institute of Technology, and Ph.D. in Aeronautics and Astronautics from Massachusetts Institute of Technology.

Timothy J. Creamer, Maj., U.S. Army. Born November 15, 1959. B.S. in Chemistry from Loyola College and M.S. in Physics from Massachusetts Institute of Technology.

Christopher J. Ferguson, Lt. Comdr., USN. Born September 1, 1961. B.S. in Mechanical Engineering from Drexel University and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Michael J. Foreman, Comdr., USN. Born March 29, 1957. B.S. in Aerospace Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Michael E. Fossum. Born December 19, 1957. B.S. in Mechanical Engineering from Texas A&M University, M.S. in Systems Engineering from the Air Force Institute of Technology, and M.S. in Physical Science from the University of Houston-Clear Lake.

Kenneth T. Ham, Lt. Comdr., USN. Born December 12, 1964. B.S. in Aerospace Engineering from the U.S. Naval Academy and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Patricia C. Hilliard. Born March 12, 1963. B.S. in Biology from Indiana University of Pennsylvania and M.D. from the Medical College of Pennsylvania.

Gregory C. Johnson. Born July 30, 1954. B.S. in Aerospace Engineering from the University of Washington.

Gregory H. Johnson, Maj., USAF. Born May 12, 1962. B.S. in Aeronautical Engineering from the U.S. Air Force Academy and M.S. in Civil Engineering from Columbia University.

Stanley G. Love. Born June 8, 1965. B.S. in Physics from Harvey Mudd College, M.S. in Astronomy from the University of Washington, and Ph.D. in Astronomy from the University of Washington.

Leland D. Melvin. Born February 15, 1964. B.S. in Chemistry from the University of Richmond and M.S. in Materials Science from the University of Virginia.

Barbara R. Morgan. Born November 28, 1951. B.S. in Biology from Stanford University.

William A. Oefelein, Lt., USN. Born March 29, 1965. B.S. in Electrical and Electronics Engineering from Oregon State University.

John D. Olivas. Born May 25, 1966. B.S. in Mechanical Engineering from the University of Texas-El Paso, M.S. in Mechanical Engineering from the University of Houston and Ph.D. in Mechanical Engineering and Materials from Rice University.

Nicholas J. M. Patrick. Born March 22, 1964. B.A.E. in Engineering from the University of Cambridge, England, M.A. in Engineering from the University of Cambridge (England), M.S. in Mechanical Engineering from Massachusetts Institute of Technology, and Ph.D. in Mechanical Engineering from Massachusetts Institute of Technology.

Alan G. Poindexter, Lt. Comdr., USN. Born November 5, 1961. B.A. in Aerospace Engineering from Georgia Institute of Technology and M.S. in Aeronautical Engineering from the U.S. Naval Postgraduate School.

Garrett E. Reisman. Born February 10, 1968. B.S. in Economics from the University of Pennsylvania, B.S. in Mechanical Engineering from the University of Pennsylvania, M.S. in Mechanical Engineering from California Institute of Technology, and Ph.D. in Mechanical Engineering from California Institute of Technology.

Steven R. Swanson. Born December 3, 1960. B.S. in Engineering Physics from the University of Colorado-Boulder, M.A.S. in Computer Systems from Florida Atlantic University, and Ph.D. in Computer Science from Texas A&M University.

Douglas H. Wheelock, Maj., U.S. Army. Born May 5, 1960. B.S. in Applied Science from the U.S. Military Academy and M.S. in Aerospace Engineering from Georgia Institute of Technology.

Sunita L. Williams, Lt. Comdr., USN. Born September 19, 1965. B.S. in Physical Science from the U.S. Naval Academy and M.S. in Engineering Management from Florida Institute of Technology.

Neil W. Woodward III, Lt., USN. Born July 26, 1962. B.S. in Physics from Massachusetts Institute of Technology and M.A. in Physics from the University of Texas-Austin.

George D. Zamka, Maj., USMC. Born June 29, 1962. B.S. in Mathematics from the U.S. Naval Academy.

The members of the 1998 rating panel and astronaut candidate selection committee were Thomas D. Akers; Joseph D. Atkinson, Jr., Ph.D.; Ellen S. Baker, M.D.; Jeffrey W. Bantle; Robert D. Cabana; Kenneth D. Cockrell; Brian Duffy; Estella H. Gillette; Linda M. Godwin; Ph.D.; James D. Halsell; Steven A. Hawley, Ph.D.; Gregory W. Hayes; Susan J. Helms; James A. Hickmon; Robert K. Holkan; Ellen Ochoa, Ph.D.; Charles J. Precourt; Duane L. Ross; John A. Rummel; Ph.D.; Loren J. Shriver; James D. Wetherbee; and John W. Young. Mr. Shriver was from NASA KSC. The others were from NASA JSC.

Group 18

NASA selected 17 candidates for the astronaut class of 2000. The group consisted of 7 pilots and 10 mission specialists, including the first candidate selected directly from the Navy's submarine community. The following astronaut candidates reported to the Johnson Space Center in August 2000 to begin extensive training in support of the Space Shuttle and Space Station programs:

Dominic A. Antonelli, Lt., USN. Born August 23, 1967 in Detroit, Michigan. B.S. in Aeronautics and Astronautics from Massachusetts Institute of Technology and M.S. in Aeronautics and Astronautics from the University of Washington.

Michael R. Barrett. Born April 16, 1959. B.S. in Zoology from the University of Washington and M.D. from Northwestern University.

Robert Behnken, Capt., USAF. Born July 28, 1970. B.S. in Mechanical Engineering and Physics from Washington University, M.S. in Mechanical Engineering from California Institute of Technology, and Ph.D. in Mechanical Engineering from California Institute of Technology.

Eric A. Boe, Maj., USAF. Born October 1, 1964. B.S. in Astronautical Engineering from the U.S. Air Force Academy and M.S. in Electrical Engineering from Georgia Institute of Technology.

Stephen G. Bowen, Lt. Comdr., USN. Born February 13, 1964. B.S. in Electrical Engineering from the U.S. Naval Academy and Degree in Ocean Engineering from Massachusetts Institute of Technology.

B. Alvin Drew, Maj., USAF. Born November 5, 1962. B.S. in Astronautical Engineering and Physics from the U.S. Air Force Academy, and M.S. in Aerospace Science from Embry-Riddle Aeronautical University.

Andrew J. Feustel. Born August 25, 1965. B.S. in Solid Earth Sciences from Purdue University, M.S. in Geophysics from Purdue University, and Ph.D. in Geological Sciences from Queen's University (Canada).

Kevin A. Ford, Lt. Col., USAF. Born July 7, 1960. B.S. in Aerospace Engineering from the University of Notre Dame, M.S. in International Relations from Troy State University, M.S. in Aerospace Engineering from the University of Florida, and Ph.D. in Astronautical Engineering from the Air Force Institute of Technology.

Ronald J. Garan, Jr., Maj., USAF. Born October 30, 1961. B.S. in Business Economics from SUNY College at Oneonta, M.S. in Aeronautical Science from Embry-Riddle Aeronautical University, and M.S. in Aerospace Engineering from the University of Florida.

Michael T. Good, Maj., USAF. Born October 13, 1962. B.S. in Aerospace Engineering from the University of Notre Dame and M.S. in Aerospace Engineering from the University of Notre Dame.

Douglas G. Hurley, Maj., USMC. Born October 21, 1966. B.S. in Civil Engineering from Tulane University.

Timothy L. Kopra, Maj., U.S. Army. Born April 9, 1963. B.S. in Computer Science from the U.S. Military Academy and M.S. in Aerospace Engineering from Georgia Institute of Technology.

K. Megan McArthur. Born August 30, 1971. B.S. in Aerospace Engineering from the University of California-Los Angeles.

Karen L. Nyberg. Born October 7, 1969. B.S. in Mechanical Engineering from the University of North Dakota, M.S. in Mechanical Engineering from the University of Texas-Austin, and Ph.D. in Mechanical Engineering from the University of Texas-Austin.

Nicole P. Stott. Born November 19, 1962. B.S. in Aeronautical Engineering from Embry-Riddle Aeronautical University and M.S. in Engineering Management from the University of Central Florida.

Terry W. Virts, Jr., Capt., USAF. Born December 1, 1967. B.S. in Mathematics from the U.S. Air Force Academy, and M.A.S. in Aeronautics from Embry-Riddle Aeronautical University.

Barry E. Wilmore, Lt. Comdr., USN. Born December 29, 1962. B.S. in Electrical Engineering from Tennessee Technological University, M.S. in Electrical Engineering from Tennessee Technological University, and M.S. in Aviation Systems from the University of Tennessee-Knoxville.

The following individuals served on the 2000 rating panel and astronaut candidate selection board: Michael P. Anderson; Ellen S. Baker, M.D.; Jeffrey W. Bantle; Robert D. Cabana; Franklin R. Chang-Diaz, Ph.D.; Kalpana Chawla, Ph.D.; Robert L. Curbeam, Jr.; Nancy J. Currie; Estella H. Gillette; Linda M. Godwin, Ph.D.; Steven A. Hawley, Ph.D.; Tamara E. Jernigan, Ph.D.; Ellen Ochoa, Ph.D.; Scott Parazynski, M.D.; William W. Parsons; Charles J. Precourt; James F. Reilly, Ph.D.; Kent V. Rominger; Duane L. Ross; John A. Rummel, Ph.D.; Loren J. Shriver; James D. Wetherbee; David R. Williams, M.D.; and John W. Young. Mr. Shriver was from NASA KSC. The others were from NASA JSC.

DUANE L. ROSS
TERESA GOMEZ
NASA Johnson Space Center
Houston, Texas

ASTRONOMY—INFRARED

Introduction

This article describes the astrophysical questions that can be addressed at infrared wavelengths, the advantages of pursuing infrared astronomy from space, the enabling technologies, and the missions that have been flown and are planned to exploit the unique potential of this wavelength range. The infrared band covers three decades—from $\sim 1\text{ }\mu\text{m}$ to $\sim 1000\text{ }\mu\text{m}$ —that encompass a very wide range of instrumental techniques and scientific issues. We have confined our discussion to rocket- and satellite-borne missions aimed primarily at astrophysical targets and have omitted the infrared instruments that have been carried on planetary probes. We draw the short-wavelength limit of our detailed discussion at $2.5\text{ }\mu\text{m}$ but note that the Near Infrared Camera and Multi-Object Spectrograph (NICMOS) instrument on the Hubble Space Telescope (HST) have operated very successfully in the 1 to $2.5\text{-}\mu\text{m}$ band; these wavelengths are also among those studied by the Diffuse Infrared Background Experiment (DIRBE) on the Cosmic Background Explorer (COBE). Wavelengths longer than $\sim 200\text{ }\mu\text{m}$ have been probed extensively from space only by the highly successful COBE spacecraft. We discuss COBE observations of the nearby Universe but cannot do justice to its extraordinary spatial and spectral measurements of the cosmologically critical cosmic microwave background radiation (CMBR).

The organization of the article is as follows. An introductory section describes the astrophysical uniqueness of the infrared and the tremendous benefits to be gained by carrying infrared instrumentation above the atmosphere. The next section discusses three key technical areas of particular importance for infrared astronomy—detectors, cryogenics, and optics—and the way they have been adapted to the space environment. Next, comes a review of previous and near-term missions for infrared astronomy from space, emphasizing the evolution of both science and technology since the first rocket experiments of the 1960s. The final mission we discuss in detail is the Space Infrared Telescope Facility (SIRTF), scheduled for launch in 2003 (1). SIRTF's advanced technology and great scientific potential will mark the end of the first phase of exploration of the Universe in the infrared band. But SIRTF is also a beginning, because it sets the stage for the missions of the next decade. We conclude with a summary of the technological challenges and scientific opportunities of these upcoming programs.

Uniqueness of Infrared

Infrared observations provide the following unique perspectives on the Universe: **The Cold Universe.** There is an inverse relationship between the temperature of an object and the peak wavelength λ of its intrinsic or blackbody radiation: $T(\lambda) = 3700/\lambda$. Here T is measured in degrees kelvin (K) above absolute zero, and λ in microns (μm). Objects with $T < T(\lambda)$ radiate very little at wavelengths less than λ . Observations at infrared wavelengths from 1–1000 μm

are thus uniquely sensitive to astronomical objects whose temperatures are from ~ 3000 K to ~ 3 K. These include the coolest stars, planets and interplanetary dust, circumstellar and interstellar matter, and, at the longest wavelengths, the Universe itself.

The Dusty Universe. Interstellar dust—microscopic particles composed of ices, minerals, and common organic and inorganic materials—is a ubiquitous constituent of astrophysical environments. The properties of this material are such that a cloud that is totally opaque in the visible or ultraviolet can be virtually transparent in the infrared; thus infrared wavelengths can probe regions—such as the core of our galaxy—which are inaccessible at shorter wavelengths. Additionally, the dust particles are heated by the shorter wavelength radiation they absorb and reradiate the absorbed power at infrared wavelengths. The majority of the radiant energy from dense, dusty regions such as star-forming clouds—and in some cases from entire galaxies—lies at infrared wavelengths because of this efficient downconversion process.

The Distant Universe. In the expanding Universe, the more distant an object is, the greater the velocity at which it recedes from us. This cosmic expansion shifts the starlight from distant galaxies into the infrared; the more distant the object, the farther out into the infrared. This expansion is characterized by the redshift parameter z : $1 + z = (\text{observed wavelength}/\text{emitted wavelength})$. The most distant known objects have $z > 6$, so that radiation from the middle of the visual band is shifted out beyond $3\text{ }\mu\text{m}$. Because $1 + z$ is also equal to the factor by which the Universe has expanded between the times of emission and absorption of the radiation, objects at $z = 5$ are seen as they were at an epoch when the Universe was only one-sixth of its present size.

The Chemical Universe. The infrared band contains the spectral signatures of a variety of atoms, molecules, ions, and solid substances—some of which will be found in any astrophysical environment. Examples range from cool ices in the interstellar medium to highly excited ions in active galactic nuclei. Infrared spectroscopy can isolate these features, determine their absolute and relative strengths, and provide an important and often unique probe of the chemical and physical conditions in these systems.

The Advantages of Space

The space environment presents powerful advantages for conducting infrared astronomical observations, which motivate the technological developments discussed below. First, in space, one is free of the absorption by Earth's atmosphere, which—even from the best mountaintop observatories—is totally opaque at wavelengths from ~ 30 to $\sim 300\text{ }\mu\text{m}$ (2). Outside of this region, there are other bands of high and moderate opacity, and atmospheric absorption remains appreciable at aircraft and balloon altitudes. Only from space do we have access to the entire infrared band. A second, equally fundamental benefit is that a space observatory is free of the blackbody radiation of Earth's atmosphere, and the space telescope can be cooled to low temperature to minimize its own blackbody radiation without fear of atmospheric condensation. Infrared observations from Earth are limited by very bright foreground radiation from the atmosphere and

the ambient temperature telescope; in space using a sufficiently cold telescope, the limiting background—set by the faint glow of the interplanetary zodiacal dust cloud—is some six orders of magnitude fainter. This is about the same factor by which the night sky at new moon is fainter than the daytime sky at high noon; note that optical astronomy is practiced at night, not during the day. The impact of this million-fold background reduction, in space, is impossible to overestimate because it produces a thousandfold increase in sensitivity, or a millionfold increase in the speed of observations. Thus the first major cryogenic infrared space observatory, the Infrared Astronomical Satellite (IRAS), revolutionized our knowledge of the infrared sky, even though it observed each point for less than about 20 seconds during its 10-month survey of the sky (3).

Complementary Approaches

Infrared astronomy is pursued very successfully from ground-, aircraft-, and balloon-borne platforms, and these sites present opportunities and capabilities complementary to the very high sensitivity and spectral access of space. The current state of the art for ground-based infrared astronomy is a series of 8- to 10-m diameter telescopes in Hawaii and Chile that provide ongoing scientific opportunities, much higher spatial resolution than achievable from space at present, and a greater variety of focal plane instrumentation—including complex spectroscopic instruments—than typically available in space observatories. Airborne observatories—exemplified by the imminent ~ 3 -m-class Stratospheric Observatory for Infrared Astronomy (SOFIA)—provide capabilities similar to those of large ground-based telescopes in most of the wavelength bands that are inaccessible from the surface of Earth. Balloon-borne instruments have been successful for specialized measurements, most notably survey observations and studies of the CMBR, and will have an important niche in the upcoming era of long-duration balloon flights.

Detectors and Detector Arrays

Modern detectors fall into two classes, bolometers and photoconductors. Bolometers are devices that change resistance when heated by absorbed radiation; they respond to radiation across a wide wavelength band, as long as they effectively absorb across this entire band. Bolometer technology has advanced dramatically in the past few years due to the application of modern semiconductor processing techniques. Photoconductors are solid-state devices in which incident photons can excite electrons from a bound state—a valence band or an impurity level—into a conduction band to produce a current in response to an applied voltage. Because the valence or impurity levels and the conduction bands are separated by a well-defined energy gap, the energy or wavelength range within which photoconductors respond to radiation is restricted. To cover a broad infrared wavelength range, a combination of different photoconductors generally needs to be employed. The current materials of choice are InSb and HgCdTe photodiodes for wavelengths shorter than $\sim 10\ \mu\text{m}$, extrinsic (doped) silicon photodetectors (Si:xx) for wavelengths from $5\text{--}40\ \mu\text{m}$, and extrinsic Germanium

(Ge:xx) for $\sim 40\text{--}200\ \mu\text{m}$. In this nomenclature “xx” denotes the specific dopant. The preferred dopants for silicon detectors are As, B, Ga, and Sb; Ga, Be, and Sb have been used as dopants for germanium. Bolometers are the detectors of choice for wavelengths longer than $200\ \mu\text{m}$ and for some applications at shorter wavelengths as well. Infrared detectors for space astronomy are discussed in detail in two recent books (4,5).

All detectors are inherently noisy. They register the incidence of arriving photons, generally referred to as “signal,” and also any number of other types of events classed as “noise.” Notable among the noise sources are thermally excited conduction electrons, or “dark current” in photoconductors, and thermal fluctuations in bolometers. Both sources of noise can be reduced by cooling the detectors to temperatures so low that thermal effects become negligible. In real-life applications, both types of detectors can be degraded by the electronics required to operate them and read them out, although modern circuit design techniques and on-chip integration generally allow minimizing these effects. In space, noise can also be generated by high-energy cosmic rays that traverse the detectors. Effective shielding against such particles becomes a high priority; in addition, special fabrication techniques can be used to reduce the susceptibility of the detectors to this ionizing radiation, as has been done with extrinsic silicon photoconductors of the type to be used on SIRTf.

A recent major advance in infrared detectors is large arrays of many active elements, or pixels, bonded to a multiplexer that is used to sample and read out the pixels. The impact of this technology on space infrared astronomy—which is similar to the CCDs used in the visible band—will be very dramatic. Used with photoconductors in the low-background space environment, the technology permits on-chip integration, so that the signal can be accumulated on the detector array and read out only when it is large enough to overcome electronic noise. Clever schemes involving multiple, nondestructive readouts of the array have been devised to suppress electronic noise further. Detector arrays are equally applicable to imaging and to spectroscopic instruments—and to photoconductor and bolometer technology, and they will certainly be used very extensively, if not exclusively, for space infrared astronomy in the future. There will continue to be a push for larger format arrays, and the next generation of infrared space experiments should use arrays of at least 1024×1024 pixels, a substantial advance over the 256×256 pixel arrays to be used on SIRTf.

The ultimate performance goal for detectors for infrared astronomy is that they permit “background-limited” observations, that is, the intrinsic detector and electronic noise should be less than the noise due to the statistical fluctuations in the rate of arrival of photons from ambient and astrophysical backgrounds. Modern infrared detectors achieve this readily in ground-based applications where the warm telescope and emissive atmosphere produce very high backgrounds. For space applications using cryogenic telescopes, the infrared background is that due to the zodiacal dust within the solar system, which is at least a million times fainter than the ground-based foreground sky. Achieving background-limited performance in this environment is quite challenging, even with the benefits of on-chip integration; for example, observations in the 3 to $5\ \mu\text{m}$ window require that dark current and electronic noise contribute less (often much less) than the equivalent of one electron/second/pixel (6). Detector

technologists and astronomers working together have responded to these challenges and improved the performance of infrared detectors by many orders of magnitude in the past two decades. The improvement has come from reducing the noise and also by improving the “quantum efficiency”—the fraction of incident photons that is absorbed by the detector. As a result, the arrays to be used on SIRTf will achieve background-limited performance for both photometry and low-resolution spectroscopy at all wavelengths.

Cryogenics

Space infrared telescopes invariably require efficient cooling or cryogenic systems; the telescope and the surrounding structure are cooled to reduce their background radiation, and the detectors are cooled to reduce their intrinsic noise and increase their sensitivity. In many applications, these effects together require cooling below 10 K. In many of the rocket and satellite instruments built to date, the entire telescope has been cooled to temperatures as low as 2 K, where no part of the apparatus emits as much radiation as the 2.73 K CMBR. This has been done by placing the entire telescope structure in direct physical or thermal contact with a pumped-liquid-helium bath, and the vacuum pump is the natural vacuum in space. The IRAS and Infrared Space Observatory (ISO) systems used this architecture, as shown in Fig. 1 (left). Liquid helium is required to achieve temperatures below ~ 5 K, but other stored cryogens that provide more cooling power per unit mass are used in applications where higher temperatures are acceptable. In practice, this is equivalent to reducing the long-wavelength limit of the instrument. Other cryogens that have been used—and the approximate temperature that they provide—are solid hydrogen (8 K), liquid neon (30 K), solid nitrogen (50 K), and liquid nitrogen (75 K).

A primary design problem for the cryogenic engineer is to minimize the heat load on cooled surfaces of the apparatus. A first step is to shield these surfaces from the principal heat source, which is solar radiation, and to use suitable combinations of low- and high-emissivity materials to reduce heat transfer within the satellite. A next step is to blanket the container that holds the cryogen with dozens of layers of aluminized mylar loosely packed within a vacuum jacket to isolate the entire system from its ambient-temperature surroundings before launch. Such a vacuum-packed cryogenic system is referred to as a dewar, named for the nineteenth-century Scottish scientist J. Dewar. A well-designed cryogenic system also uses the cold effluent gas generated as the cryogen evaporates or sublimates to cool the surrounding structures and further reduce the heat load on the cryogen.

A second design challenge is the construction of apparatus sufficiently sturdy to survive launch and also to maintain optical alignment between the cooled telescope and the guide telescopes or gyroscopic components that typically operate at ambient temperature within the spacecraft bus. The mechanical rigidity required tends to go hand in hand with high thermal conductivity, which the design must avoid; this requires using low-thermal-conductivity materials that have high strength, such as epoxy-glass composites.

Considerable attention has also focused on minimizing heat loads on the cryogen by passively radiating intercepted heat into cold space; this is referred to

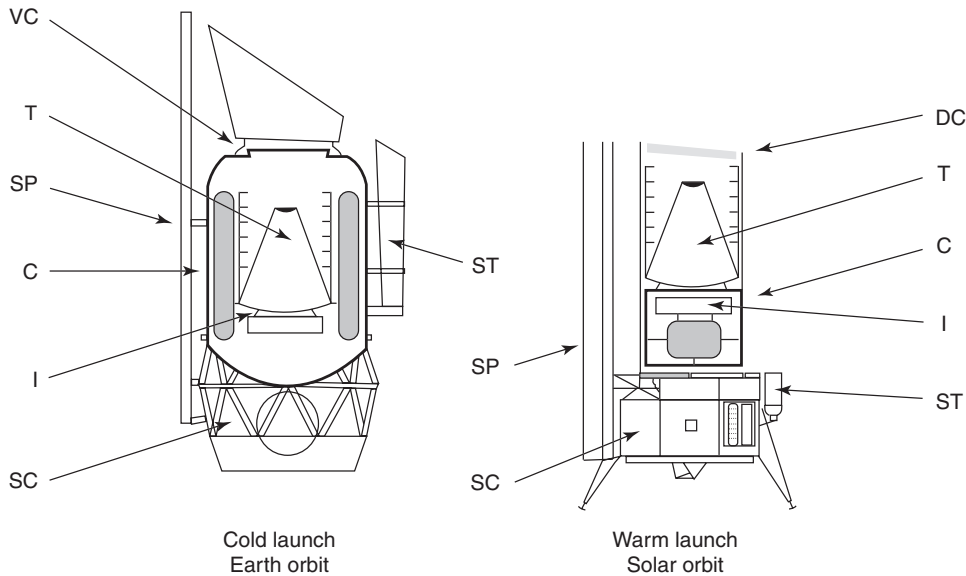


Figure 1. This figure compares the cold launch architecture used for the ISO and IRAS observatories (left) with the warm launch architecture to be used for SIRTf (right). Each is shown in cutaway view. Certain components, such as the spacecraft (SC), the solar panel (SP), and the startracker (ST) are common to both systems. In addition, the telescope (T) and instrument package (I) are identical in size for the two. Each also includes a cryostat (C), containing the liquid helium cryogen in a separate helium tank, which is shown shaded. In the cold launch system, the telescope is located within the cryostat and cooled by direct contact with the cryogen tank. In the warm launch system, the cryostat and cryogen tank can be much smaller, and the telescope is cooled by conduction and by the cold boil-off helium gas. This architecture works in the solar orbit because the cylindrical thermal shields that surround the telescope cool radiatively to 40 K or below, so there is very little parasitic heat diffusing inward toward the telescope. The cryostat must withstand atmospheric pressure, and it is much larger and more massive for the cold launch than for the warm launch system. In the former, it surrounds the entire telescope and supports a heavy vacuum cover (VC). In the warm launch system, the telescope is launched at ambient temperature and pressure, protected only by a lightweight dust cover (DC). The sawed-off conical sunshade at the top of the cryostat is required in an Earth-orbiting system by the Sun–Earth-orbit geometry. A much smaller sunshade is needed in the solar orbit system because Earth is not a concern.

as “radiative cooling” (7). The SIRTf telescope, described later, exploits the favorable thermal environment of its heliocentric orbit by using a hybrid cryogenic system in which the instruments and detectors are cryogenically cooled, whereas the telescope is launched warm and is cooled by a combination of radiation, conduction, and effluent cryogen (Fig. 1 right). This approach has many advantages over that used in earlier missions such as IRAS and ISO, in which the entire telescope was placed within the cryostat. It leads to a lower mass cryogenic system for a fixed telescope size and decouples the size of the telescope from that of the cryostat. Thus this hybrid approach is certain to be adopted for large infrared telescopes in the future.

In a more extreme application of radiative cooling, it may be possible to cool the entire telescope to a temperature acceptable for many purposes without

using cryogenics. Just how low a temperature can be reached in practice is still not clear, but many designers now assume that equilibrium temperatures as low as 30 K could be within reach at 1 astronomical unit (the radius of Earth's orbit) from the sun. Lower temperatures might be achieved by a telescope operating in the outer solar system. At such low temperatures, a well-designed telescope that has exceptionally low-emissivity mirrors might radiate at such low levels in the wavelength range shorter than 100 μm that the primary and secondary mirrors require no active cooling at all. Active cooling would be required primarily for the detector arrays and their immediate housings. The reduced cooling requirements of such a system might be satisfactorily met by an acceptable, though still substantial, charge of cryogen, or by closed-cycle refrigerators required only to pump heat at low rates.

A stored cryogenic system always has a limited lifetime: unless replenished (an approach which has not been adopted for any astronomical mission), the cryogen eventually is fully depleted, the system warms up, and the mission comes to an end. To increase mission life spans, a variety of recyclable cryocoolers—both mechanical and electrochemical—have been under intense study (8). In principle, they could extend lifetimes indefinitely. In practice, a nagging long-term problem has been the limited reliability of closed-cycle, low-temperature refrigerators designed to operate in the vacuum of space. In the laboratory, such systems have often failed catastrophically after only a few months. No refrigerator of this type has ever operated in the laboratory continuously for a 10-year span. Yet this is the expected mission lifetime of many infrared astronomical space facilities now on the drawing boards. In 1998, NASA successfully tested a particular type of mechanical cooler on a Shuttle mission—a reverse Brayton-cycle cryocooler that can cool detectors to temperatures as low as 60–70 K. This test showed that operation under weightless conditions was not a problem for this type of cooler, but long-term reliability is still an open question, though the same coolers have a good record in the laboratory and run reliably for many months to a few years. Reliable closed-cycle cryocoolers are certain to affect critically the design and life spans of future infrared astronomical missions in space. The reverse Brayton-cycle cryocooler described before has been retrofitted to the NICMOS instrument on HST to extend its useful lifetime beyond the almost 2 years achieved with the initial charge of solid nitrogen.

Some types of highly sensitive infrared detectors now being planned for future missions operate effectively only at temperatures in the millikelvin range. Additional cooling beyond that achievable with liquid ^4He must be provided for them. In the laboratory, a variety of techniques has already been developed to reach such low temperatures; often, they require a succession of stages that might employ combinations of thermoelectric, liquid ^3He , $^3\text{He}/^4\text{He}$ dilution, adiabatic demagnetization, or other refrigerators. For long-duration astronomical space missions, reliable refrigerators will be required to provide these low temperatures continuously or cyclically. Again, these devices are often used in tandem; on the ESA/NASA Planck mission to study the CMBR, a hydrogen sorption refrigeration provides an 18-K heat station for a mechanical cooler which, in turn, provides a 4.5 K stage for a dilution refrigerator that cools the bolometer detectors to ~ 100 mK (9).

Light Collectors

With few exceptions, light collectors for the infrared are all-reflecting telescopes whose optical components may be aluminized, or gold-coated, depending on the wavelength range. Conventional telescopes image a portion of the sky onto a focal plane to provide accurate maps. Occasionally, however, the astronomer is interested in diffuse radiation that is not localized but arrives from all over the sky. For such observations, a carefully designed horn, an all-reflecting funnel, is generally employed to gather radiation from a large but well-defined field of view in the sky onto the smallest possible detector. These two types of light collectors were used, respectively, on the ISO/IRAS and COBE spacecraft.

For space applications, there is a premium on lightweight optics because the mass of the entire satellite scales with the mass of the optical system it must support, and, in turn, a more massive satellite requires a larger and more expensive launch vehicle. For an infrared mission, there is the added complication of the increased thermal conductivity of the beefier structure required to support the more massive optical system. Both IRAS and SIRTf, as discussed later, used all-beryllium optical systems because of the favorable strength-to-mass ratio of this material. The 85-cm diameter SIRTf primary, for example, has a mass of 15 kg and an areal density of 26 kg/m^2 . By comparison, the Hubble Space Telescope primary mirror has an areal density of 180 kg/m^2 . As telescope apertures beyond $\sim 4 \text{ m}$ diameter are considered for future missions, another launch vehicle limitation, set by the physical size of the payload shroud, is encountered. Thus planning for the 8-m diameter Next Generation Space Telescope (NGST) is based on ultralightweight panels of glass, beryllium, or composite materials, whose areal density is no greater than 15 kg/m^2 . It would deploy after launch to achieve the desired aperture (6).

Filters

Filters isolate wavelength ranges of particular interest to the astronomer. For imaging and photometry, a well-defined, broad wavelength range needs to be isolated. Carefully designed transmission filters are usually used for this purpose. For spectroscopy, different types of spectrometers that select numerous narrow-wavelength intervals are inserted between the light collector and the detector or detector arrays. The most common types of spectrometers for infrared are prism or grating “dispersive systems” that separate out radiation directionally, according to wavelength, and interferometers. Fabry–Perot interferometers select one narrow-wavelength range at a time; Michelson and other two-beam, multiplex interferometers transmit many wavelengths simultaneously but have to be swept through a range of settings to encode unambiguously and register the flux detected at each wavelength. Later, we describe spectrometers by their spectral resolving power R , defined as $\lambda/\delta\lambda$, where λ is the operating wavelength and $\delta\lambda$ is the finest discernible spectral detail. The unique requirements placed on filters and spectrometers for space applications are largely environmental and have to do with surviving launch or the ionizing radiation in space, or achieving low mass or volume, rather than with the device’s functionality or performance.

Early Rocket Instrumentation

In the mid-1960s, a collaborative effort between Cornell University and the Naval Research Laboratory led to the design of liquid-nitrogen-cooled and eventually liquid-helium-cooled, rocket-borne infrared telescopes. Early Cornell designs incorporated a parabolic primary mirror with an 18-cm aperture and focal ratio length ratio $f/0.9$. The entire telescope, except for the entrance aperture, was surrounded by the liquid. Four different types of detectors were flown on many of these flights to sample the spectral range from $5\text{ }\mu\text{m}$ to 1.6 mm (10). Using this apparatus, the total flux in a field of view roughly 1° in diameter was first successfully measured for the galactic center and four other regions in the central portions of the Milky Way, at 5, 13, 20, and $100\text{ }\mu\text{m}$. A first spectral measure of the radiation emitted by the solar system's zodiacal dust was also obtained (11).

Some years later, results from a survey conducted in a series of rocket flights were published by the U.S. Air Force Cambridge Research Laboratories (now the Air Force Geophysical Laboratories). The group initially flew liquid-neon-cooled telescopes that had 10-cm apertures and detectors sensitive to radiation at $12\text{--}14\text{ }\mu\text{m}$. Each of six detectors in a linear array surveyed a $10' \times 10'$ field of view. Later, the group began all-sky surveys using satellite-borne instrumentation and also began observations across wider spectral ranges. Early results of one of these surveys at 4.2, 11.0, 19.8, and $27\text{ }\mu\text{m}$ were cataloged and published by Price and Walker (12).

Although rockets have not been extensively used for infrared astronomy in recent years, large-format infrared detector arrays may enable significant science in the limited duration of a rocket flight. For example, a 16.5-cm rocket-borne telescope instrumented with a 256×256 InSb array and cooled by supercritical helium has been flown to search for a faint halo of low-mass stars enveloping a nearby edge-on spiral galaxy (13).

The Infrared Astronomical Satellite

The first true infrared survey of the sky from space was carried out by the Infrared Astronomical Satellite (IRAS), jointly sponsored by the United States, the Netherlands, and Great Britain (3). Approximately two-thirds of the 300-day mission that lasted from January to November 1983 was devoted to an unbiased survey of the sky that succeeded in charting 98% of the celestial sphere in four broad wavelength bands. IRAS was launched into a polar orbit at the day–night terminator which precessed about 1° per day. In this “Sun-synchronous” orbit the Earth/Sun/spacecraft geometry varied only slowly, so that the survey could be executed by a simple scanning strategy. Observations were carried out with an all-beryllium 57-cm aperture, $f/9.6$ Richey–Chrétien telescope whose focal plane was cooled to 3 K and featured a total of ~ 60 Si:As, Si:Sb, and Ge:Ga discrete photoconductors; each had a separate JFET amplifier readout. The detectors covered, respectively, the 12-, 25-, 60-, and $100\text{-}\mu\text{m}$ bands, using Ge:Ga appropriately filtered to cover the last two. A low-resolution spectrometer covered the wavelength range from $7.5\text{--}23\text{ }\mu\text{m}$.

A measure of the mission's success was the cataloging of some 250,000 celestial sources; the vast majority had never before been detected in the infrared. No area of modern astrophysics was untouched by IRAS. A few of the many scientific highlights include

1. The discovery of galaxies that emit up to fifty times more energy at far-infrared wavelengths than in the optical domain and also emit from 100 to 1000 times as much total power as our own galaxy, the Milky Way. The existence of such highly luminous infrared galaxies came as a huge surprise.
2. The discovery of disks composed of fine dust grains that orbit around a number of stars that, in many ways, were reminiscent of our own Sun. This dust, it was conjectured, is the remnant of an originally far more massive circumstellar cloud of gas and dust from which a system of planets had already formed and initiated further astronomical searches for signs of planets around these stars. A sharply defined, though far fainter, set of dust rings was also found orbiting our own Sun, as were enduring trails of dust left by the passage of solar system comets.
3. The successful measurement of spectra for planetary nebulae and a variety of other sources at wavelengths previously inaccessible due to telluric absorption. IRAS also identified patchy infrared emission from the diffuse interstellar medium, referred to as "infrared cirrus" because of its similarity to the thin, streaky clouds in Earth's atmosphere. Infrared cirrus is important as a tracer of matter within our galaxy and as a potential source of interference in observations of distant galaxies.

The Cosmic Background Explorer, COBE

COBE, built by NASA and launched into a polar orbit identical to that of IRAS in 1989, was dedicated to the study of the microwave and infrared background radiation in space (14). It carried three instruments; two of them, the Diffuse Microwave Radiometer (DMR) and the Far Infrared Absolute Spectrophotometer (FIRAS) were used to study the CMBR—the isotropic blackbody radiation whose temperature is ~ 2.73 K and is believed to be a relic of the Big Bang in which the Universe was born. The third experiment, the Diffuse Infrared Background Experiment (DIRBE), measured the background at infrared wavelengths from 1–200 μm . DIRBE and FIRAS were cooled by liquid helium and mapped the entire sky repeatedly during the ~ 10 -month cryogenic lifetime of COBE. The critical components of DMR were cooled passively to ~ 140 K; this instrument operated for about 4 years.

The CMBR carries important cosmological information, and DMR and FIRAS were extremely successful, respectively, in measuring the spatial structure in that radiative field and in establishing its blackbody nature at a very high degree of precision. These important cosmological experiments will not be discussed further here. Of course, both FIRAS and DMR also measured the foreground radiation from our own galaxy. FIRAS was a polarizing Michelson interferometer instrumented with helium-cooled bolometers as detectors. It

obtained spectra of the galactic emission from $\sim 100\text{ }\mu\text{m}$ to $\sim 3\text{ mm}$ using a 7° field of view. Of particular interest was its detection of emission from oxygen, carbon, nitrogen, and carbon monoxide from gas in the Galaxy (15).

DIRBE made measurements in ~ 15 wavelength bands, using a variety of discrete photodiodes and photoconductors from $1\text{--}100\text{ }\mu\text{m}$ and helium-cooled bolometers at 140 and $240\text{ }\mu\text{m}$. All measurements were referenced to an internal cold, black reference surface so that the absolute sky brightness was determined.

The principal scientific results from DIRBE include (16)

1. An improved determination of the distribution of the infrared radiation from the zodiacal dust cloud within the solar system, which has led to improved models of the dust cloud and its infrared emission. DIRBE also confirmed IRAS' discovery of a modest enhancement of emission in Earth-trailing direction, which is attributed to temporary gravitational trapping by Earth of zodiacal dust particles that are spiraling inward towards the Sun.
2. Measurements of the large-scale distribution of infrared radiation from the Galaxy, including both the far infrared radiation from $25\text{--}200\text{ }\mu\text{m}$ that samples the distribution of heated dust, and the near-infrared radiation from $1\text{--}25\text{ }\mu\text{m}$ that is indicative of the large-scale distribution of stars in the Galaxy.
3. Detection of an isotropic background of infrared radiation at 140 and $240\text{ }\mu\text{m}$ that arises from outside the Galaxy and may be attributable to the integrated effects of star-forming galaxies at redshifts $z \sim 1$ to 2 .

The Infrared Space Observatory (ISO)

The Infrared Space Observatory, built and launched by the European Space Agency (ESA), was the first comprehensive infrared astronomical space observatory. NASA and the Japanese Space Agency (ISAS), provided important technical, operational, and scientific support. ISO mapped celestial sources and analyzed them through spectroscopy, photometry, and linear polarization studies (17).

On the night of 16–17 November 1995, an Ariane 4 rocket launched ISO into a highly elliptical, 24-hour, circumterrestrial orbit, where the observatory operated with great success until its helium ran out and instruments began warming up in April 1998. The spacecraft in orbit was 5.3 m long, 2.3 m wide, and its mass was approximately 2500 kg . At launch, it carried a superfluid helium charge of 2300 liters, which maintained the Ritchey–Chrétien telescope, the scientific instruments, and the optical baffles at temperatures of $2\text{--}8\text{ K}$. The diameter of the telescope's fused silica primary mirror was 60 cm . A three-axis-stabilization system provided an absolute pointing accuracy of a few seconds of arc and stability of a fraction of an arc second in both jitter and long-term drift. The telescope was diffraction-limited down to wavelengths of roughly $5\text{ }\mu\text{m}$. Four instruments formed the core of the scientific payload:

1. A camera containing two 32×32 pixel arrays: InSb for the wavelength range $2.5\text{--}5.5\text{ }\mu\text{m}$, and Si:Ga for the range $4\text{--}18\text{ }\mu\text{m}$. Each array could be

operated with a selection of filters for broadband spectrophotometry or continuously variable filters (CVF) for low-resolution ($R \sim 40$) imaging spectroscopy and could view sources through three linear polarizers oriented relative to each other at angles of 60° .

2. A photometer covered the entire wavelength range from 2.5–240 μm . It employed Si:Ga detectors that gave peak response at 15 μm , Si:B detectors that gave peak response at 25 μm , unstressed Ge:Ga detectors that gave peak response at 100 μm , and stressed Ge:Ga detectors that gave peak response at 180 μm . Stressed detectors are mounted in a miniature clamp or vise that applies high mechanical pressure to the crystal, thereby extending their wavelength response. At 100 and 200 μm , the instrument housed, respectively, 3×3 and 2×2 arrays of unstressed and stressed Ge:Ga to facilitate mapping. Multiple apertures, multiple filters, and polarizers were used for photometric and photopolarimetric measurements in each range. Scanning and mapping operations were carried out at all wavelengths. Two grating spectrophotometers, each with a 64-element linear Si:Ga detector array, provided spectra with resolving power $R \sim 100$ at 2.5–5 and 6–12 μm .
3. A short-wavelength spectrometer included both grating and Fabry–Perot (FP) instruments. Grating spectra were available for the entire wavelength range from 2.38 to 45.2 μm , with resolving power $R \sim 1000$ –2000. The FP mode covered the 11.4- to 44.5- μm range and gave resolving power of a factor of 20 higher. For the grating mode, the detectors were InSb at 2.38–4.08 μm , Si:Ga at 4.08–29 μm , and Ge:Be at 29–45.2 μm . For the FP mode, Si:Sb was used out to 26 μm , and Ge:Be from 26–44.5 μm .
4. A long-wavelength spectrometer provided coverage from 43–196.9 μm . A grating provided resolving power $R \sim 150$ –200. A FP mode permitted observations at $R \sim 6800$ –9700. The 10 detectors, arranged in a linear array on a curved surface, were Ge:Be at 43–50 μm , unstressed Ge:Ga at 50–100 μm , and stressed Ge:Ga beyond 110 μm .

Among the scientific highlights of ISO were

1. The detection of water vapor throughout the interstellar medium of the Galaxy. Before ISO, the infrared emission from interstellar water vapor could not be detected because telluric water vapor absorbs at precisely the emission wavelengths. Water vapor, however, can be one of the primary coolants of interstellar clouds, and the extent of this cooling needed to be understood to assess the extent to which it facilitates protostellar collapse.
2. Detection of polycyclic aromatic hydrocarbons in the spectra of galaxies. These large molecules were well known in our galaxy, but ISO had the sensitivity needed to show that their emission dominates the 5- to 12- μm emission from nearby spiral galaxies as well. The emission from these molecules is due to radiative fluorescence: a molecule is excited by optical or ultraviolet radiation, quickly rereadiates the energy of a single absorbed photon, in the infrared and returns to the ground state.

3. Inventories of extragalactic source-counts at wavelengths ranging from 4–175 μm . These are of particular value in understanding the origins of the extragalactic diffuse infrared radiation detected by the COBE mission. Many of the randomly observed galaxies appear to be ultraluminous, indicating that they contain substantial regions of massive star formation or that they harbor an active galactic nucleus that possibly surrounds a massive central black hole.

The Space Infrared Telescope Facility (SIRTF)

NASA is developing SIRTF for launch in 2003; it has a projected cryogenic lifetime greater than 5 years (1). SIRTF will be an observatory for infrared astronomy from space and will complete NASA's family of Great Observatories (the other members are the Hubble Space Telescope, the Compton Gamma Ray Observatory, and the Chandra X-ray Observatory).

SIRTF culminates the four decades of technology development and scientific progress described above. SIRTF will be the first space mission to use exclusively the imaging and spectroscopic power of large format infrared detector arrays. SIRTF's all-beryllium telescope that incorporates an 85-cm diameter primary mirror and is diffraction-limited down to 6.5 μm , defines the state of the art for ultralightweight cryogenic optics. The SIRTF telescope and cryogenic system will be carried on a fairly standard spacecraft bus that provides pointing control, power, data storage, and communication. The pointing system is built around an external autonomous star tracker that controls and reports the spacecraft orientation at more than 2'' accuracy and has a reaction wheel/gyro control system. Visible light sensors in the cold focal plane can sense stars simultaneously using the external star tracker to track the relative orientation of the telescope and star tracker lines of sight. This pointing system architecture was used on ISO as well. The SIRTF spacecraft also incorporates a nitrogen gas system that is used to unload the reaction wheels if they accumulate too much angular momentum; the magnetic torquer bars used for this function in Earth-orbiting spacecraft would not work on SIRTF because it is far outside Earth's magnetosphere.

Unlike the missions previously described, all of which operated in Earth orbit, SIRTF will be placed into an Earth-trailing heliocentric orbit, drifting slowly away to reach a distance of ~ 0.5 AU from Earth after 5 years. In this orbit, SIRTF is free of the heat load from Earth and provides good access to the sky for target selection and scheduling. SIRTF will launch with the telescope warm and the instruments at helium temperature. In space, the telescope and its surrounding thermal shields cool radiatively to ~ 40 K, and the effluent helium from the cryogenic tank cools the telescope down to its operating temperature of ~ 5.5 K (see Fig. 1b). The thermal shields remain at 40 K and below throughout the mission, so that the parasitic heat conducted into the telescope is very small. As a result, the heat load that dissipates the SIRTF cryogen and determines the lifetime of the mission comes largely from the focal plane instruments.

This hybrid, radiative, cryogenic cooling system is facilitated because SIRTF can maintain an attitude in its solar orbit in which the solar panel is always

oriented toward the Sun and shades the thermal shields that control the telescope temperature. These structures, in turn, are optimized to minimize the heat transferred from the solar panel and to radiate to space any heat that is transferred. This approach would not work in near-Earth orbit because the heat load from Earth would occasionally be incident on the thermal shields and turn the radiator into an absorber.

This optimized cryogenic system, together with the low-power dissipation of its instruments and the elimination of parasitic heat loads, makes SIRTf a much more efficient system cryogenically than any of its predecessors. SIRTf carries 350 liters of helium at launch, and a lifetime of $5 +$ years is predicted, based on an average instrument power dissipation of ~ 5 mW. By comparison, ISO was launched with ~ 2300 liters of helium and achieved a lifetime of ~ 2.5 years with an average instrument power dissipation of ~ 10 mW.

SIRTf will carry three array-based focal plane instruments:

1. A near-infrared camera that provides imaging simultaneously in four bands at 3.6, 4.5, 5.8, and $8\text{ }\mu\text{m}$. Both of the 3.6- and $5.8\text{-}\mu\text{m}$ channels image the same field of view in the sky; this is made possible by a dichroic filter that transmits $5.8\text{ }\mu\text{m}$ and reflects $3.6\text{ }\mu\text{m}$. An adjacent field of view is imaged at 4.6 and $8\text{ }\mu\text{m}$ in a similar fashion. Each band uses a 256×256 pixel array hybridized to a 256×256 MOSFET multiplexer. The detector material is InSb in the 3.6- and $4.5\text{-}\mu\text{m}$ bands and Si:As in the 5.8- and $8\text{-}\mu\text{m}$ bands.
2. A spectrometer that provides low resolving power ($R \sim 60\text{--}120$) spectroscopy from $5\text{--}40\text{ }\mu\text{m}$ and higher resolution spectroscopy ($R \sim 600$) from $10\text{--}38\text{ }\mu\text{m}$. The spectrometer consists of four physically distinct modules; each contains a 128×128 array (Si:Ga for the shorter wavelengths, Si:Sb for the longer wavelengths) illuminated by an optical train of mirrors and gratings. The use of detector arrays allows these modules to be very compact and efficient and obviates the need for moving parts. The higher resolution modules use two diffraction gratings so that, an entire octave of the spectrum can be cross-dispersed across the entire array and measured simultaneously. In the lower resolution modules, a long entrance slit is used to permit obtaining spectra simultaneously at many spatial points. A portion of the array in one of these modules is also used for the precision target acquisition required to place a source on a narrow spectrograph slit, thereby alleviating the absolute pointing requirements placed on the spacecraft.
3. An imager/photometer that provides imaging and low-resolution spectrophotometry at wavelengths between 25 and $160\text{ }\mu\text{m}$. This instrument uses a 128×128 Si:Ga array at $25\text{ }\mu\text{m}$ but incorporates two Ge:Ga arrays for longer wavelength measurements. The 32×32 Ge:Ga array used by SIRTf at $70\text{ }\mu\text{m}$ is composed of eight 4×32 submodules; each in turn consists of four 1×32 linear arrays, coupled to a 1×32 amplifier/multiplexer. The 2×20 array used at $160\text{ }\mu\text{m}$ is similarly built up of four 2×5 pixel modules with the added complication that the module construction allows for the mechanical stress needed to extend the long-wavelength response of Ge:Ga from 120 to beyond $160\text{ }\mu\text{m}$. These arrays represent substantial

advances in the state of the art and point the way toward still larger arrays for future applications. Note that although these arrays have fewer pixels than those described before, they are to be used at longer wavelengths where the diffraction-limited image size is larger compared to the field of view. Thus they provide comparable sampling of the telescope's focal plane; in fact, all three arrays are designed to sample the image fully to allow numerical postprocessing of the data to enhance the spatial resolution.

The scientific return of SIRTf cannot be forecast because its capabilities represent such a great advance beyond what has been possible in the past and also because the bulk of the observing time on SIRTf will be dedicated to programs to be proposed and carried out by the general scientific community. However, based on the science return from the other missions described before, we anticipate that SIRTf will lead to great advances in our understanding of such problems as

1. the formation and early evolution of galaxies, stars, and planets;
2. the physical processes that power the objects of highest luminosity in the Universe;
3. the chemical composition of interstellar and circumstellar matter;
4. the nature of the coolest, lowest luminosity stars and star-like objects in the solar neighborhood; and
5. the properties and interrelationships of comets, asteroids, interplanetary dust, and other small bodies in the solar system.

In addition, SIRTf's large arrays, very high sensitivity, and long lifetime give this mission great potential for discovering new phenomena.

Other Infrared Missions Already Flown

Other significant infrared space astronomy missions are described briefly here:

1. The Spacelab II Infrared Telescope—1985. A 15-cm diameter helium-cooled telescope was flown on Spacelab-2 and made infrared measurements between 2 and 120 μm . It provided data about the structure of the Galaxy (18) and about the infrared background environment on the Space Shuttle.
2. The Midcourse Space Experiment (MSX)—1995. MSX carried a 35-cm aperture off-axis telescope and five linear Si:As arrays that mapped the sky in a push-broom fashion in bands from 4–22 μm . Although it was primarily designed to scan Earth's limb, it carried out a number of astrophysical experiments and produced excellent images of the entire galactic plane at ~ 18 arcsec resolution (19).
3. The Infrared Telescope In Space (IRTS)—1995. IRTS was an ISAS program with significant NASA participation. IRTS had a 15-cm diameter liquid-helium-cooled telescope and four varied focal plane instruments covering wavelengths from 3–800 μm . It was carried on a Japanese satellite

called the Space Flyer Unit and surveyed $\sim 6\%$ of the sky in a 5-week lifetime (20).

4. The Near Infrared Camera and Multi-Object Spectrograph (NICMOS)—1996. NICMOS is a replacement focal plane instrument, which was installed on HST. It was instrumented with three 256×256 HgCdTe arrays that carried a range of filters that covered the 1- to $2.5\text{-}\mu\text{m}$ spectral band and was optimized for high spatial resolution imaging. NICMOS was cooled by solid nitrogen and achieved a lifetime of slightly less than 2 years. This was somewhat less than expected because of a partial failure of the cryogenic system in orbit (21).
5. The Submillimeter Wave Astronomy Satellite (SWAS)—1998. SWAS was the first space mission to carry radio-type (heterodyne) receivers for spectroscopic exploration. SWAS has a 55×71 cm near-optical quality off-axis primary mirror and two heterodyne radiometers with Schottky barrier diode mixers and a single acousto-optical spectrometer. SWAS is surveying the galactic plane in the emission of atomic carbon, molecular oxygen, water vapor, and carbon monoxide in five transitions between 538 and $615\text{ }\mu\text{m}$ (22).

Future Missions

A number of missions are planned or proposed for the next two decades to go beyond even the great scientific and technical accomplishments described before. These include

1. The Infrared Imaging Surveyor (IRIS) is a Japanese mission that will employ a 70-cm telescope cooled to 6 K, using a hybrid cryogenic system that incorporates both liquid helium and mechanical coolers to achieve a lifetime in excess of 1 year. Its primary mission is to conduct an all-sky survey at wavelengths of $50\text{--}200\text{ }\mu\text{m}$ at an angular resolution of $30\text{--}50$ arc seconds. The detectors to be used for this purpose are stressed and unstressed Ge:Ga. IRIS is likely to be launched in the first half of the decade 2000–2010 (23).
2. The Far Infrared Space Telescope (FIRST) is a mission sponsored by the European Space Agency with substantial participation by NASA. Its primary mission is to provide detailed spectroscopy and imaging for the $80\text{--}670\text{-}\mu\text{m}$ spectral range. It features a 3.5-m passively cooled, primary mirror that illuminates three different liquid-helium-cooled instruments. FIRST is to be launched by an Ariane 5 rocket into a Lagrangian point L2 orbit in the second half of the decade 2000–2010 (24). FIRST will share its launch vehicle with the Planck mission to study the CMBR mentioned earlier. FIRST has been renamed Herschel by ESA.
3. The Next Generation Space Telescope (NGST) is an international mission led by NASA, that will use a radiatively cooled, ultralightweight 8-m diameter telescope that will deploy following launch (6). A suite of focal plane instruments using very large detector arrays will observe at wavelengths

$\sim 0.6\text{ }\mu\text{m}$ to beyond $10\text{ }\mu\text{m}$. NGST will greatly extend the scientific results of both HST and SIRTf in this wavelength region. NGST will be launched by an Atlas rocket into a Lagrangian point L2 orbit toward the end of the decade 2000–2010 and will have a scientific lifetime greater than 5 years.

4. Interferometers in Space (25). All of the missions described before were built around a single telescope. To achieve much higher resolution than the $\sim 0.1\text{--}1\text{ arcsec}$ achievable with an $\sim 8\text{-m}$ telescope, it will be necessary to use the techniques of interferometry, in which infrared radiation collected by two widely spaced telescopes can be brought together to achieve angular resolution comparable to that which would be provided by a single telescope whose aperture is equal to the separation of the two telescopes. Ultimately, this technique will be employed in NASA's Terrestrial Planet Finder, scheduled to launch in the 2010–2015 time frame to image Earth-like planets around nearby stars.

ACKNOWLEDGMENTS

Portions of this work were carried out at the Jet Propulsion Laboratory, California Institute of Technology, operated under a contract with the National Aeronautics and Space Administration.

BIBLIOGRAPHY

1. Fanson, J.L., et al. *Proc. SPIE* 3356: 478 (1998).
2. Traub, W., and M. Stier. *Appl. Opt.* 15: 364 (1976).
3. Neugebauer, G., et al. *Astrophys. J.* 278: L1 (1984).
4. Rieke, G., and K. Visnovsky. *Detection of Light: From the Ultraviolet to the Submillimeter*. Cambridge University Press, Cambridge, 1994.
5. McLean, I.S. *Electronic Imaging in Astronomy*. Wiley, New York, 1997.
6. Mather, J.C., et al. In E.P. Smith, A. Koratkar (eds). *Science with the NGST*, Astron. Soc. Pacific Conf. Series, 133: 3 (1998).
7. Hawarden, T.G., et al. In S.J. Bell Burnell, J.K. Davies, and R.S. Stone (eds). *Next Generation Infrared Space Observatory*. Kluwer, Dordrecht, 1992, pp. 113–144.
8. Glaiser, D.S., et al. In R.G. Ross, Jr. (ed.), *Cryocoolers 10*. Kluwer Academic/Plenum, New York, 1999, pp. 1–19.
9. Collaudin, B., and T. Passvogel. *Cryogenics* 39: 157 (1999).
10. Harwit, M., J.R. Houck, and K. Fuhrmann. *Appl. Opt.* 8: 473 (1969).
11. Houck, J.R., et al. *Astrophys. J.* 169: L31 (1969).
12. Price, S.D., and R.G. Walker. *The AFGL Four Color IR Sky Survey: Catalog of Observations at 4.2, 11.0, 19.8, and 27 μm* . AFGL-TR-76-0208, 1976.
13. Uemizu, K., et al. *Astrophys. J.* 506: L15 (1998).
14. Boggess, N.W., et al. *Astrophys. J.* 397: 420 (1992).
15. Bennett, C.L., et al. *Astrophys. J.* 434: 587 (1994).
16. Hauser, M.G., et al. *Astrophys. J.* 508: 25 (1998).
17. Kessler, M.F., et al. *Astron. Astrophys.* 315: L27 (1996).
18. Kent, S.M., et al. *Astrophys. J. Supp.* 78: 403 (1992).
19. Carey, S.J., et al. *Astrophys. J.* 508: 721 (1998).
20. Murakami, H., et al. *Pub. Astron. Soc. Jpn.* 48: 41 (1996).

21. Thompson, R.I., et al. *Astrophys. J.* 492: L95 (1998).
22. Melnick, G.J. *Proc. SPIE* 3357: 348 (1998).
23. Murakami, H., et al. *Proc. SPIE* 3356: 471 (1998).
24. Pilbratt, G.L., et al. *Proc. SPIE* 3356: 452 (1998).
25. Unwin, S., R. Stachnik (eds). *Working on the Fringe: An International Conference on Optical and IR Interferometry from Ground and Space*. To be published in Astron. Soc. Pacific Conf. Series, 2000.

MICHAEL WERNER
Jet Propulsion Laboratory
Pasadena, California

MARTIN HARWIT
Cornell University
Ithaca, New York

B

BIOLOGICAL RESPONSES AND ADAPTATION TO SPACEFLIGHT: LIVING IN SPACE—AN INTERNATIONAL ENTERPRISE

Introduction

Attempting to predict human responses to space travel, physicians and researchers of the 1950s speculated that microgravity and spaceflight itself would present significant challenges—if not barriers—to the human body (1,6). They hypothesized that the combined stresses of launch acceleration, weightlessness, radiation, and heavy deceleration upon reentry would be incapacitating. At the very least, they predicted that the bodily systems sensitive to gravity-based cues would function improperly or not at all. Given this grim forecast, the initial focus was to demonstrate that life, away from Earth, could survive space travel and subsequent return to Earth's gravity. Faced with this challenge, both the United States and Soviet Union turned first to ground simulations, such as immobilization studies, and then to the study of animal test subjects launched on board high-altitude balloons, suborbital, and orbital rockets. What followed is a series of biological satellites that carried a variety of living specimens, from isolated cell cultures to whole instrumented organisms. The Soviet Union relied primarily on canines to provide such data, whereas the United States chose primates for such experimentation.

There is not much doubt that in the beginning of the U.S. space exploration effort, the primary motivation was very real competition with the Soviets. What is remarkable is that in 1969, at the very height of this competition, when the

Americans first landed on the Moon, a long-lasting collaborative program in space biology was initiated by the United States and the Soviet Union. In that year, NASA flew the U.S. built "Biosatellite". This was a complex satellite that was placed in near-earth orbit with an appropriately instrumented chimpanzee as the payload. "Biosatellite" was flown and was successfully recovered after a week in orbit. Numerous measurements were made, but ultimately NASA rated the mission as only "partially successful," because of the death of the primate early in the postflight period, due to complications from loss of fluids and potentially an infection, thus loss of valuable data.

The "Biosatellite" mission was managed by the NASA-Ames Research Center because Ames was NASA's lead center in space biology. With the Apollo program in full gear, NASA decided to cancel all planned biological flight programs, except for a small experiment involving the behavior of a frog's otolith in zero gravity. The response from the leadership at Ames was to seek permission from NASA Headquarters to initiate collaboration with the Soviets in this area. Drs. Harold P. Klein and Joseph C. Sharp succeeded in making the case, and they received permission to go ahead. Eventually, several distinguished Soviet scientists were invited to visit Ames. This group included Professor Alexander Oparin, who was a distinguished expert in exobiology, and Academician Oleg Gazenko, an important leader in the Soviet space organization.

Eventually, an agreement was reached that would call for U.S. scientists to develop biological payloads to be flown on Soviet "Vostok" spacecraft. Dr. Sharp was the coordinator on the U.S. side, and Dr. Eugene Ilyin, Dr. Gazenko's deputy for biology, was the Soviet leader of the program. The first payload was flown in 1975. Since then, there has been an average of one flight every 2 years, and some significant scientific results were obtained early in the program (2,4). The early success of this program led to a sustained cooperative effort that survived several U.S. – Soviet confrontations during the Cold War. The Soviet invasion of Afghanistan in late 1979 caused President Carter to terminate many relationships that had been established as the Cold War wound down. For example, U.S. participation in the Moscow Olympic Games in 1980 was canceled. However, early in 1980, an American biological payload was flown aboard a Soviet satellite. Appeals from the scientists involved in this work prevailed to keep things going.

The success of the U.S. – Soviet collaboration, which also included exchange of biomedical data from piloted missions, was followed by the much higher profile 1975 Apollo. Soyuz Test Project Joint U.S. – Soviet flight and docking of the respective countries, spacecraft in low Earth orbit. This program was the result of a diplomatic effort during the Nixon – Ford Administrations. To promote a "détente" in the Cold War, it was felt that a joint U.S. – Soviet space mission would be helpful. The joint U.S. – Soviet crew performed a number of scientific experiments, but it was clear that the primary justification was symbolic. In that sense, the linkup of the U.S. Apollo and the Soviet Soyuz spacecrafts, which led to the famous "handshake in space," was a successful effort that aided the process of "détente" (3).

Serious thinking about possible scientific U.S. – Soviet collaboration of people in space was triggered by the successful first flight of the new Space Shuttle "Columbia" in April 1981. The Soviets had deployed the "Salyut 7" space station in 1982, and the existence of these capabilities once again led to the development

of some specific proposals for collaborative efforts. In 1982 and 1983, Air Force Lieutenant General James A. Abrahamson, who was serving as NASA's Associate Administrator for Space Flight at the time, proposed a "people exchange" between "Columbia" and "Salyut 7." The idea would be to have "Columbia" fly close to the "Salyut 7" station. An Astronaut from "Columbia" would perform an extravehicular activity (EVA), do a "space walk," and enter "Salyut 7." This would be followed by a Soviet Cosmonaut leaving "Salyut 7" and joining the crew of "Columbia." A preliminary proposal was developed but was not implemented for various technical and political reasons that existed at the time.

The end of the Cold War in 1991 led to an expansion of U.S. – Russian collaborative activities in space. A series of meetings between the U.S. Vice President, Albert V. Gore, and Russian Prime Minister Victor Chernomyrdin resulted in agreements that involved the collaborative use of the Russian "Mir" Space Station, as the Russian participation in and contribution to the International Space Station Program. James Beggs and Dr. Hans Mark, the NASA Administrator and Deputy Administrator, secured Administration and Congressional support to initiate the latter facility in 1984. From the very beginning, the U.S.-led space station program was intended as an international effort. The European, Canadian, and Japanese Space Agencies had strong roles in the beginning, and in 1993, Russia was brought into the program. The Russian contribution was to build various segments of the Space Station and to use Russian "Proton" launch vehicles to resupply the Space Station. The Russians also agreed to provide the Soyuz TM spacecraft as an initial crew rescue vehicle. The NASA Administrator Daniel S. Goldin initiated the International Space Station redesign effort to accommodate all of the international partners, improve research capacity, and provide long-duration spaceflight training for U.S. astronauts and ground personnel. It was also a rehearsal for international partners of living and working together in space. In addition, NASA resources were used for more than 300 Russian space scientists for competitively selected research projects and collaboration with their U.S. counterparts. This effort provided much needed resources to the Russian space science community at the time of economic transition in the post-Soviet era. The Research program on the MIR Space Station was developed by Drs. Arnauld Nicogossian and Carolyn Huntoon for NASA. Dr. Nicogossian was responsible for overseeing the implementation of the research program on "MIR" and the funding of the Russian space science community.

These efforts began in 1995 with a rendezvous in space between the Space Shuttle "Discovery" and the "Mir" Space Station. An astronaut-cosmonaut exchange was executed later in 1995 in a docking of Space Shuttle "Atlantis" and "Mir" that finally fulfilled what General Abrahamson had proposed more than a decade earlier. The difference between what was done in 1995 and what had been proposed is that no EVA was necessary to do the 1995 mission. An adapter had been constructed to make it possible for the U.S. Space Shuttle to dock directly with the "Mir" Space Station. The first Shuttle – Mir mission was followed by a number of docking missions involving both American astronauts and Russian cosmonauts. In this way, several Americans experienced truly long-duration spaceflights for the first time.

With the advent of the International Space Station, U.S. – Russian collaboration quickened. It is now quite routine to have “mixed” U.S. – Russian crews working for long periods of time on the International Space Station. Thus, in the span of four decades, sustaining life in space has evolved considerably from the first flights intended to prove that humans could endure microgravity. Since the first forays into space by Gagarin, Shepard, Grissom, Titov, and Glenn, the United States has explored the Moon, the Soviet Union/Russia has maintained a series of space stations in orbit for more than 25 years, and American astronauts have shepherded the orbital Shuttle through more than 100 missions and hundreds of sophisticated biomedical experiments. One of those missions carried Senator John Glenn, on his second flight 36 years after the Mercury 6 orbital flight, intended to underscore another NASA – NIH cooperation in the study of aging. An intrinsic, even critical, component of this evolution has been to define and overcome the biomedical challenges of human space flight and also to understand the role of gravity in critical life processes. After all, life originated and evolved under the constant pull of Earth’s gravity, using this force from cradle to grave in ways that are still poorly understood. Only recently, pulling together ground and flight data obtained from integrated and interdisciplinary biological and clinical experiments, scientists have begun to unravel the effects of gravity on living systems. This understanding was made possible by the collaboration of scientists around the world and the participation of the National Institutes of Health. The ability of gravity to up- or down-regulate certain genes involved in musculoskeletal metabolism and the capacity of the nervous system to adapt its function rapidly as a result of changing gravitational forces are just two of the most fascinating discoveries reported recently (4).

What follows is an attempt to present a concise overview of life in space, physiological responses to this new environment, and associated health implications for future space travelers. Consideration must be given to the hostile, and yet dynamic environment of space; to the craft that protects the crew members from the harsh environment, allowing them to navigate in space; and to life itself as it adjusts to this novel environment. Today we know that some adaptive changes, such as immune and hormonal responses, are primarily the response to the stresses of confinement, isolation, and spacecraft design, not necessarily to the unique effects of the space environment.

The Spaceflight Environment

The first astronauts and cosmonauts were required to function—eat, drink, communicate, and move—for extended periods of time in a novel and complex environment. Early biomedical studies demonstrated that the combined factors of spaceflight did exact a toll on the physical performance and health of crew members. Furthermore, early crew members and aerospace engineers realized that spaceflight was really the sum of several complex factors, only one of which is microgravity. The success of each mission is driven by a number of parameters—crew performance, crew health, the internal spacecraft environment, system performance, and the external environment—whose sum represents a challenge to safety and, most importantly, human survival.

External. In orbit, a spacecraft moves around Earth in a constant state of free fall that produces microgravity. Because all organisms on Earth have evolved and developed in the presence of gravity, the absence of this force imparts adaptive changes that are initiated immediately upon exposure. Some of these continue throughout the course of the mission. The net results of these changes are altered physiological function and structure that can offset health and performance in space and postflight.

Although the most conspicuous characteristic of the space flight environment is reduced gravity, a number of other factors contribute to its biomedical effects on humans. Primarily, space is a hostile environment to life. It is distinguished by profound fluctuations in temperature ranging from 220 K in the stratosphere to 1000 K in the thermosphere (5), the lack of a breathable atmosphere due to a near-perfect vacuum, and a number of constant and intermittent radiative events. In addition, the spacecraft is usually subjected to micrometeoroid bombardment and recently, to a large amount of human-made debris. All of these external and internal environmental events require constant monitoring to protect crew health and mission safety (6,7).

The radiative environment is comprised of several sources of ionizing radiation, a general term that encompasses particles that can alter molecular electrons upon contact: solar energetic particles emitted during solar flares, particles trapped in Earth's magnetic field, and galactic cosmic radiation (8). Although each of these sources consists of various types of space radiative particles, the most significant barriers to mitigating radiative exposure are

1. an incomplete understanding of the detrimental effects caused by ionizing radiation, and
2. an inability to predict and model radiative events fully in time to protect a space-faring crew or biologically based life support system (9).

Presently, crew and spacecraft interior are monitored by a series of passive and active dosimeters, and ground-based monitoring can warn of impending solar flares so that the mission can be abbreviated. The Soviet Union and NASA performed a number of experiments, using biosatellites, the Space Shuttle, the Mir Space Station, and now the International Space Station (ISS), to understand the effects of protons, neutrons, electrons, and heavy cosmic ions on the function of living organisms. Now, it is postulated that one of the effects of continuous exposure to galactic – cosmic radiation can result in cellular genomic instability, manifested as generational cancer and/or malformations. Comparison of the measurements made during the Skylab mission in 1971–73 to those obtained from the NASA-MIR (ISS Phase I) experiments showed that the South Atlantic Anomaly, and the associated radiation, has moved westward and north, reflecting a displacement in Earth's electromagnetic fields (4,9). NASA, together with the ISS partners, is now beginning to standardize radiative measurements and data analysis. Since 1998, NASA and the National Institutes of Health developed a research program to study the long-term genetic implications of exposures to constant low-level radiation. One of the most unexpected findings from the NASA – Russian cooperation was the realization that the spacecraft's aluminum

shell results in fragmentation and transformation of incoming energetic rays (bremsstrahlung) that contribute up to 30% of the total space radiative exposures (10).

Internal. Not all biomedical changes or medical events observed in spaceflight result from altered gravity or the influences of the external environment. The human-rated spacecraft must protect the crew from the hostile external environment and provide the resources necessary to support human life and work (see U.S. Manned Spaceflight: Mercury to the Shuttle). The design and performance of such spacecraft has evolved considerably from the first space capsules to the “shirt-sleeve” environment of the International Space Station. Common to all spacecraft environments, however, is the requirement for appropriate air, water, temperature, and pressure and consideration for the human factor in design and performances, that is, the environment must be precisely monitored and maintained using a minimum of resources including power, mass, and crew time (11).

Although the spacecraft itself must be rigorously designed to shield crew members from both constant and intermittent ionizing radiation, cumulative exposure levels within the spacecraft must be constantly monitored and assessed. Currently, both active and passive dosimeters of different sensitivities are used to track the exposure of each individual and selected regions in the spacecraft. As missions of increasing duration are emphasized, however, monitoring requirements will become more stringent, and highly accurate modeling and prediction will become a central means of protecting crew members. Because missions beyond earth’s orbit cannot be rescheduled or aborted to avoid predicted radiative events, exploratory missions will require a heavily shielded shelter, preferably constructed with a nonmetallic and high hydrogen content material within the spacecraft, or some other solution for protecting space-faring crews from high radiative exposures.

The spacecraft environment itself provides additional challenges to crew health and safety. A majority of these constraints results from the combined physical environment of spaceflight. Crew members are expected to live in confinement far distant from friends and family are subjected to extreme scrutiny and pressure to complete their work in a timely and consistent manner (12). This stress and isolation is compounded by the fact that a sunrise or sunset occurs every 90 minutes in low Earth orbit. Although initially disconcerting to crew members, altered dark–light cycles have a significant physiological effect on the quality and quantity of sleep and ultimately, performance. The sum of these challenges, both mental and physical, can exact a considerable toll on crew members unless they have appropriate support from ground personnel, sufficient personal time and space, and flexible work/rest schedules.

Numerous analog environments, including extended bed rest and Antarctic wintering-over expeditions, have been explored to understand the cumulative effects on human performance and psychosocial health. Human experimentation in space is a tedious and difficult task: differences among missions, variability in environmental parameters, small sample size, and mission constraints have made the design and execution of controlled experiments quite challenging (6). The Soviet Union/Russia operated its spacecraft at near 760 torr and normal atmospheric gas composition and pressure, whereas the United States adopted a

one-gas (100% oxygen) one-third atmospheric (250 torr) pressure. The intent was to simplify the life support system and minimize decompression sickness during space walks. This approach was used for early human missions and was intended to save time in the race to the Moon. At the conclusion of the Apollo missions, NASA engineers were faced with a prospect of oxygen toxicity during the long-duration Skylab missions. NASA and U.S. Air Force life scientists solved this problem by adding 20% nitrogen to Skylab's atmosphere. In the meantime, the tragic death of the three Soviet cosmonauts returning from the first Space Station, Salyut 1, created concerns in NASA regarding the safety of long-duration biomedical research on Skylab. Following a cooperative agreement signed in 1971, Soviet life scientists shared the medical findings from the Salyut 1 mission with their NASA counterparts, demonstrating that rapid decompression, due to a hatch seal failure, resulted in crew death. This mutual collaboration ensured confidence and the success of Skylab missions, and it also generated an unprecedented and very successful scientific and professional medical relationship between the physicians and scientists of the two countries. This relationship persists to this day, despite many political and management changes in both countries. The major engineering and biomedical challenge to both countries was presented during the planning and execution of the Apollo-Soyuz Test Project, docking two spacecraft that had markedly dissimilar atmospheres, posing a potential fire hazard to the Soyuz Spacecraft, and the risk of bends to the crews. A special transfer airlock module was developed for this purpose. This module was carried by the Apollo spacecraft and docked with the Soyuz; during the joint phase, the atmospheric pressure in the Soyuz was reduced by 155 torr, and the oxygen concentration was slightly increased. During a 3-day period, the crews visited each other's spacecraft three times; no bends occurred despite repeated recompressions and decompressions in the airlock.

The Space Shuttle presented a real challenge for space walks and the risk of bends. NASA adopted a standard atmosphere composition and pressure, in a decision made in response to the demands from the international scientific community, but the space suit pressure and gas composition changed little. NASA's budgetary constraints precluded the opportunity of using advanced space suits designed by the Ames Research Center. This concern for bends continues to persist in the ISS era.

Over the years, life scientists learned that an answer to the challenges posed by the engineering design, mission constraints, and uncertainties of research in space is to use multiple species, in addition to humans, to document similarities in responses to experimental variables among species. This requires integrating many different biological specimens into major life science experiments, especially during the Shuttle/Spacelab era. The necessity to execute integrated international experiments to increase the scientific return was demonstrated by the U.S. and Soviet/Russian life science cooperation (4,10).

To date, more than 350 individual biological experiments were flown in space. NASA launched 20 suborbital and 4 orbital missions, and Soviet Union/Russia launched more than 22 spacecraft (Biosattelite) dedicated primarily to fundamental biology. The last 11 Biosatellites included international participation, with major contributions from the NASA Ames Research Center. A total of 35 different species were flown by NASA, Russia, Europe, Japan, Ukraine,

and Canada. The results of these investigations are presented in the following sections.

Biomedical Challenges

The organism uses a complex set of biological tools to sense, process, and respond to the ever-changing environment, be it the normal habitat on Earth or the close quarters of spacecraft. Full understanding of the relationships between these various tools is yet to be achieved. Table 1 presents concisely the sum of our current knowledge in this area.

Because of the complexity and individual nature of human adaptation to spaceflight, changes are often considered on a systemic basis. Although this approach is useful for understanding system-specific functional alterations, it does not fully characterize the effects of living and working in the spaceflight environment. As a result, more recent studies have focused on whether crew members maintain appropriate levels of functional performance, the ability to perform key activities such as intra- and extravehicular activity (IVA, EVA) or emergency egress during or following long-duration spaceflight. Aerobic capacity, a measure of the amount of oxygen consumed during a single bout of maximal exercise, reflects the integrated performance of the cardiovascular, nervous, and musculoskeletal systems. As illustrated in Fig. 1, aerobic capacity is diminished by spaceflight but does show a relatively rapid return toward preflight levels upon return to Earth (6,13).

Cardiovascular Deconditioning. The cardiovascular system has evolved in the presence of gravity as an intricate network of vasculature that contains blood and is powered by the heart. This vasculature is composed of both muscular arterial vessels that supply oxygenated blood to tissue and nonmuscular venous vessels that return blood to the heart. Baroreceptors and stretch-sensitive receptors monitor the critical parameter of blood pressure in vessels throughout the body and adapt to postural changes. On Earth, simple motions such as sitting, standing, or reclining result in significant and rapid responses to changes in gravitational force imposed on the body.

Spaceflight presents a challenge to the cardiovascular system that is generally stabilized by the fifth week of flight. In even longer Soviet space flights (3 months to 1 year or more), a slight increase in heart rate has been noted, particularly toward the end of the mission (7,14). Nevertheless, cardiovascular deconditioning, by Earth's standards, appears to be a self-limiting phenomenon on that does not worsen with flight duration (unless other medical conditions, such as dehydration or infection prevail) and does improve upon return to Earth. In space, cardiovascular responses represent an appropriate adjustment to a new environment, in which the gravitational load placed on the heart is considerably less than on Earth. Fluid pooling no longer occurs in the lower extremities but is instead localized to the upper body. Physically, this shift is revealed by facial edema, sinus congestion, decreased calf girth, and leg volume ("bird legs"). This shift is perceived as excess fluid, which in turn affects a series of immediate but long-lasting changes. An immediate decrease in plasma volume occurs, in addition to a more gradual loss of approximately 10% of red blood cell

Table 1. A Summary of the Physiological Changes Associated with Human Spaceflight

Physiological measure	Change associated with short spaceflights (<3 weeks) ^a	Change associated with long spaceflights (>3 weeks) ^a
Cardiovascular System		
Resting heart rate	<i>Peaks during launch and entry; ↓ in-flight; ↑ postflight</i>	<i>In-flight normal or slight ↑; ↓ orthostatic tolerance postflight</i>
Resting blood pressure	↔	Diastolic blood pressure ↔ or ↓
Total peripheral resistance	↓; ↔ at landing despite ↓ stroke volume and ↑ heart rate	Tendency toward ↓
Stroke volume	↓ postflight	↓ postflight
Exercise capacity	↔ or ↓ ≤ 12% after flight; increased HR for same O ₂ consumption; efficiency ↔	Submaximal exercise capacity ↔
Body Fluids		
Total body water	3% ↓ by flight day 4 or 5	↑ (first in-flight sample) followed by slow ↓
Plasma volume	↓ in-flight and postflight	↓ approximately 15% during first 2–3 weeks; recovery begins after 60 days and is independent of mission duration
Hemoglobin	↔ or slightly ↑ after flight	
Red blood cell (RBC) mass	↓ after flight (approximately 9%); sustained at least 2 weeks and associated with decreased erythropoietin level	
Plasma lipids	↓ cholesterol and triglycerides in-flight	
Plasma glucose	↓ during and immediately after flight	↓ first 2 months, then leveled off
Serum/plasma electrolytes	↑ K and Ca in-flight; ↓ Na in-flight; ↓ K and Mg postflight	↓ Na, Cl, and osmolality; slight ↑ K and PO ₄
Urine electrolytes	<i>Postflight ↑ in Ca, creatinine, PO₄ and osmolality; ↓ in Na, K, Cl, and Mg; decreased citrate</i>	↑ osmolality, Na, K, Cl, Mg, Ca, and PO ₄ ; ↓ uric acid excretion; decreased citrate
Insulin		↓
Urine volume	Postflight ↓	↓ early in flight

Sensory Systems		
Gustation and olfaction		
Vision	Subjective and varied experience; no impairments noted Intraocular tension ↑ in-flight and ↓ at landing; postflight ↓ in visual field, visual motor task performance and contrast discrimination; ↔ in-flight contrast discrimination or distant and near visual acuity; dark-adapted crews reported light flashes with eyes open or closed; retinal blood vessels constricted postflight 40–70% astronauts/cosmonauts exhibit in-flight neurovestibular effects, including immediate reflexive motor responses and space motion sickness; motion sickness symptoms appear early in flight and subside or disappear in 2–7 days. <i>Postflight ataxia and disrupted postural control</i>	Same as shorter missions Light flashes reported by dark-adapted subjects; frequency related to latitude In-flight vestibular disturbances are the same as those for shorter missions; cosmonauts have reported occasional reappearance of illusions during long-duration missions. <i>Ataxia postflight and significant alteration in postural control mechanism</i>
Musculoskeletal System		
Height	Slight ↑ (~1.3 cm/0.5 in) during first week in flight, and 1 day recovery to baseline Postflight weight ↓ average 3.4%; about two-thirds is due to water loss and the remainder due to lean body mass and fat ↓ 15% on flight day 2	↑ during first 2 weeks in flight by a maximum of 3–6 cm (1.2–2.4 in); stabilizes thereafter In-flight weight losses average 3–4% during the first 5 days and are probably due to loss of fluids; thereafter, weight ↑ or ↓ for the remainder of the mission and is related to metabolism
Extracellular fluid volume	↓ postflight	Center of mass shifts headward
Total body volume	↓ during and after flight, with 1–2 weeks recovery to baseline	
Muscle strength	Postflight EMGs from gastrocnemius suggest ↑ susceptibility to fatigue and ↓ muscular efficiency; EMGs from arm muscle ↔	
EMG analysis		
Bone density		<i>Os calcis density ↓ postflight; radius and ulna show variable changes, depending upon method, ↓ bone density in lumbar vertebrae, pelvis and femoral trochanters</i>
Calcium balance	↑ negative Ca balance in flight	<i>Excretion of Ca ↑ during first month of flight; fecal Ca excretion ↓ until day 10 then ↑ continually throughout mission; Ca balance becomes ↑ negative as the mission progresses</i>

^a ↔ unchanged; ↓ decrease; ↑ increase; N/A: not measured.

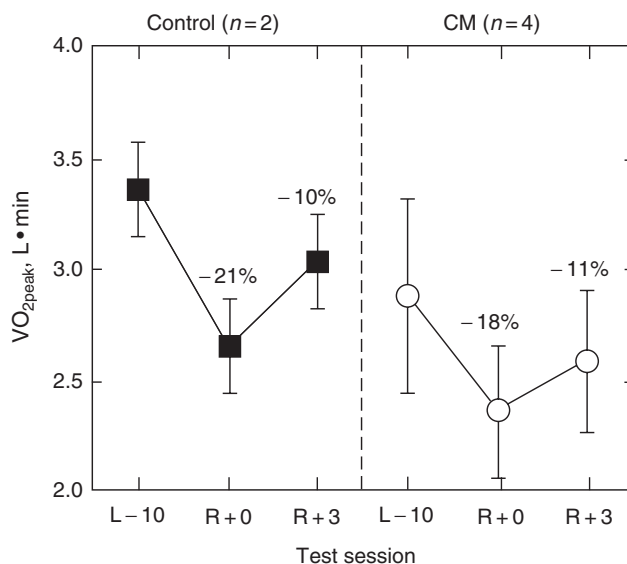


Figure 1. Aerobic capacity of crew members before launch (L – 10 days) and after flight (Return, Day 0 and Return, Day 3) as measured by maximum O₂ uptake in liters per minute during an exercise stress test. Control subjects did not perform any exercise, whereas CM subjects exercised using an onboard cycle ergometer up to 48 hours before landing (13).

mass (15). This condition is primarily attributed to a decrease in circulating blood volume. Systematic investigations have proven difficult because of individual differences in diet, sleep patterns, exercise, medications, and fluid intake associated with various space missions. Recent studies have focused on secretion of hormones (such as norepinephrine), baroreceptor changes with time, and the role of the central nervous system in regulating of the cardiovascular system (16).

Microgravity-induced cardiovascular adaptation becomes a medical problem only after crew members are subjected to accelerative forces during reentry or upon return to the constant 1-g stress on Earth. As early as the American Gemini program, cardiovascular deconditioning was documented in 100% of crew members. One component of this deconditioning is orthostatic intolerance, the inability to function effectively against gravitational stress, such that simple actions like sitting and standing may result in episodes of weakness, dizziness, or fainting. A standard measure of orthostatic intolerance is the stand test, in which recently returned crew members are asked to stand upright for several minutes after a period of reclining; by monitoring blood pressure and heart rates during this functional challenge, researchers can associate significantly altered arterial pressures with adaptation to space flight and the gravitational forces of landing. As shown in Fig. 2, approximately 20% of crew members showed altered levels of systolic and diastolic pressure following flight. Depending on the duration of the space flight and the amount of exercise performed in flight, the return of cardiovascular function to preflight values may take as long as 2 weeks. Routinely used countermeasures include aerobic exercise, fluid and electrolyte

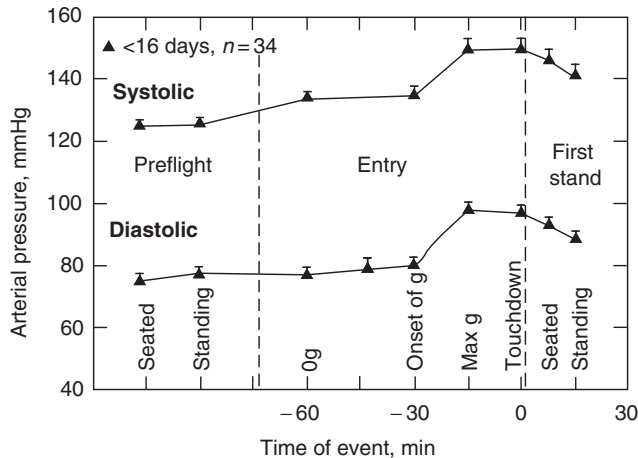


Figure 2. Systolic and diastolic pressure responses of crew members to entry, landing, and egress shows altered reactions to the standard orthostatic challenge, the stand test. Blood pressure was measured preflight while seated and standing. The middle section of this graph represents blood pressure values in-orbit (an average of 130/76 mm Hg), during reentry, and until touchdown and egress from the orbiter seat. Values while seated in the orbiter are much higher than preflight values (16).

replenishment (especially before return to Earth), and exposure to simulated gravity via the lower body negative pressure device (6,7,17,18).

Neurosensory Disturbances. The central nervous system (CNS) controls both perception of and interaction with the environment. The sensory system, including the visual, vestibular, and proprioceptive organs, responds to environmental stimuli and supplies a constant flow of input to the CNS. In conjunction with a visual image of surroundings, the vestibular and proprioceptive systems supply additional information relating to orientation, balance, and limb location. The CNS processes this information and then directs the musculoskeletal system's movement and interaction with the environment. Each step in this intricate process is contingent upon a constant inflow of information about the surrounding environment. In microgravity, however, the CNS must adapt to a loss of sensory and proprioceptive input, and it also must also respond to reduced muscular capacity, including functional and structural changes in muscle tissue. Adaptation to unexpected or even absent sensory information is neither an instantaneous nor a constant process; thus identifying the mechanisms responsible and the appropriate countermeasures is somewhat of a challenge. Neurosensory adaptations have traditionally been difficult to measure but indirectly are evidenced through in-flight and postflight changes in crew performance.

Clinically, the most important vestibular disturbance associated with spaceflight is space motion sickness (SMS). As Titov noted, most crew members do experience a sensation of bodily inversion, which soon passes but can recur due to rapid movement (19). More susceptible individuals, however, develop a full host of SMS symptoms (20). Russian and U.S. demographics suggest that SMS affects 40–70% of astronauts and cosmonauts, depending on the classification of

symptoms. SMS occurs early in the mission, typically within the first 3 days. Symptoms range from minimal discomfort to nausea and vomiting, accompanied in rare cases by pallor and sweating. Head and body movements tend to worsen the discomfort. When the symptoms are severe, crew performance can be affected and mission efficiency severely compromised. During the Apollo IX mission, for example, certain crew activities were delayed by 24 hours due to space motion sickness.

The medical basis for space motion sickness is not fully understood, partly because the phenomenon can be studied effectively only during spaceflight. Guedry et al. (21) summarized studies of motion sickness in flight and on the ground, in which the most plausible explanation for neurosensory changes is the "sensory conflict" hypothesis. According to this concept, the usual sensory inputs to the vestibular receptors of the inner ear are no longer present in microgravity, causing altered processing of sensory information and ultimately resulting in altered motor responses. A series of elegant biological experiments conducted aboard the dedicated Spacelab Life Sciences Missions 1 and 2 and the NASA-NIH Neurolab Mission, dedicated to the Decade of the Brain, demonstrated the plasticity of the CNS in response to altered gravitational forces. The CNS does develop new coping strategies in novel environments, which facilitates orientation and navigation. Of interest is the fact that individuals who use internal cues for orientation fare better in the SMS manifestation and adaptation process than those use external cues. Repeated exposures to the space environment reduce the severity of SMS and the length of time required to adapt. Anecdotal data suggests that women are less susceptible to SMS. Finally, Earth-based simulators cannot identify individuals susceptible to SMS. Absorption of orally ingested medications is altered in space, and drugs are not effective in preventing or treating SMS. Injectable PhenerganTM has been successfully used to treat SMS in-flight and has minimal side effects (6).

Both the Russian and American human spaceflight programs recognize the complex interactions within and between the sensory-motor, nervous, cardiovascular, and muscular systems (10). A significant finding from postflight research is the ataxia present during ambulation in the majority of flown individuals. Current countermeasures against this ataxia involve resistive exercise combined with compressions along the longitudinal axis of the body.

Musculoskeletal Alterations. An integrated response from the skeletal, muscular, connective tissue, and nervous systems permits movement in a 1-g environment. This response is predicated on the fact that certain directional forces must be overcome to complete ordinary tasks, such as lifting an object or walking down stairs. In the microgravity environment, however, these directional forces are altered; the result is a cascade of functional and structural changes in the physiological systems that control locomotor tasks. Collectively, these changes yield reductions in strength, power, and endurance that ultimately influence crew-members' ability to perform routine motor activities (Table 2). The changes are ordinarily indicated in flight by a progressive decrease in total body mass, leg volume, and muscular strength. As weight-bearing muscles and bones adapt to the microgravity environment, several symptoms are manifested. Disturbances in postural and motor coordination, locomotion function, and equilibrium can be seen, and alterations in proprioceptor activity and spinal reflex

Table 2. Mean Strength Performance of Skeletal Muscle on Landing versus Preflight (n = 17) during Concentric and Eccentric (Extension) Motions of Selected Muscle Groups (13) Pre>Landing (p<0.05)

Muscle group	Test mode	
	Concentric	Eccentric
Back	- 23 (\pm 4)*	- 14 (\pm 4)*
Abdomen	- 10 (\pm 2)*	- 8 (\pm 2)*
Quadriceps	- 12 (\pm 3)*	- 7 (\pm 3)
Hamstrings	- 6 (\pm 3)	- 1 (\pm 0)
Tibialis Anterior	- 8 (\pm 4)	- 1 (\pm 2)
Gastroc/Soleus	1 (\pm 3)	2 (\pm 4)
Deltoids	1 (\pm 5)	- 2 (\pm 2)
Pecs/Lats	0 (\pm 5)	- 6 (\pm 2)*
Biceps	6 (\pm 6)	1 (\pm 2)
Triceps	0 (\pm 2)	8 (\pm 6)

mechanisms occur. Although all of these changes appear to be dependent, at least to some extent, on flight duration, they have been reversible, and no adverse sequelae have been reported thus far.

A primary indicator of changes in bone and muscle is body mass: in-flight weight losses of 3–4% were seen in early, short-duration spaceflights. With the advent of longer missions, most weight loss took place during the first three to five flight days, and a much more gradual decline thereafter (22). The findings suggest that a significant part of the initial change in body mass is due to the loss of fluids, either through diuresis or decreased thirst and fluid intake (18), and that subsequent losses are due to metabolic imbalances and/or muscle atrophy. These changes appear to be self-limiting, the largest weight losses recorded (6 to 7 kg) are independent of mission duration. In more recent long-duration space missions, where adequate caloric intake and physical exercise have been maintained by some crew members, actual weight gains have been reported. Such weight gains probably reflect an overall increase in fatty tissue, which was more than sufficient to offset losses of muscle tissue (7,10). In any event, body mass lost in flight is rapidly regained in the postflight period.

Muscle Atrophy. Muscle atrophy results from structural and functional changes in muscles. These changes are most readily apparent in the postural or antigravity muscles, such as the gastrocnemius abdominal, back, and neck muscles. Skeletal muscles exhibit numerous alterations in strength and endurance properties, including force- and power-generating capacities, shortening and relaxing rates, neural activation patterns, protein expression, and metabolic utilization profiles. Concomitant with these muscular changes, connective tissues undergo similar atrophy and functional alteration. At the molecular level, both slow-twitch and fast-twitch muscle fibers are affected. The process of functional and structural change is progressive and can be controlled to some extent by increasing caloric intake, dietary adjustments, and intensive strength exercises.

Evidence of the deterioration of muscle during spaceflight comes from several sources. In-flight measurements of leg volume (Fig. 3) show an initial rapid

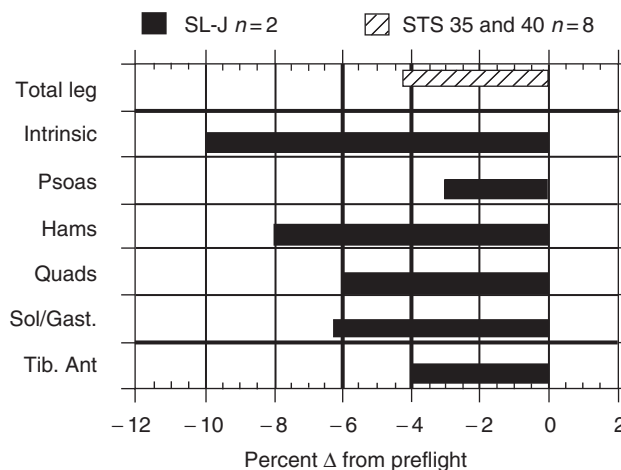


Figure 3. Percent decreases in volumes of various leg muscles, measured by magnetic resonance imaging, from three extended duration Orbiter Shuttle missions (13).

decrease that can be attributed to the headward fluid shift and is followed by gradual recovery. Postflight biostereometric measurements of Skylab astronauts demonstrated more general losses of volume from the abdomen downward, although losses in the abdomen and buttocks were attributed to the loss of fat (23). Postflight urinary analyses reveal in-flight increases in the excretion of a number of metabolites associated with muscle breakdown, such as nitrogen, potassium, creatine, and amino acids. Metabolic balance studies and electromyographic analyses of muscular activity further substantiate the deterioration of muscle function during spaceflight. Electromicroscopic analysis of human and rodent muscle biopsies showed decreased production of slow-twitch myosin fibers (endurance) and normal distribution of fast-twitch myosin fibers (dexterity). Additional investigations performed on rodents flown on Spacelab missions demonstrated the selective effects of gravity on the type of muscle myosine production (4).

Bone and Mineral Changes. Removing muscular forces and weight from bones, as occurs in bed rest or having a limb in a cast, causes a loss of bone mineralization, known as disuse osteoporosis. During space flight, crew members experience a form of musculoskeletal disuse in which levels of bone mineral are decreased. Early studies of bone mineral changes using X-ray densitometry suggest that large amounts of bone may be lost during relatively brief periods of spaceflight, and countermeasures to this loss are mandatory for long-duration missions (24). The 12 crew members who participated in the Gemini 4, 5, and 7 and Apollo 7 and 8 missions averaged 3.2% postflight losses of bone density from the calcaneus (heel bone) compared with preflight baseline values. Some losses were also observed from the radius and ulna after these early flights. Data from Soviet/Russian cosmonauts and U.S. astronauts who flew on the Mir station show a continued loss of 1% of bone mass per month, even using exercise as a countermeasure. Resistive exercise, hormone therapy, and drugs such as bisphosphonates are currently being evaluated as potential countermeasures to bone loss.

In sum, these changes to the locomotion system means that crew members are at risk for increased falls, bone fractures, and limited mobility—conditions, which at a minimum, could make emergency egress a challenge.

Immunologic Alterations. The immune system defends the body against any cell, substance, or organism not recognized as self. As such, it is affected by both environmental and physiological fluctuations that occur during spaceflight. Although results from some studies are contradictory, most generally recognize an increase in the immune cells responsible for the immune response, known as leukocytes. More specifically, changes in the leukocyte population, particularly in the relative percentage of T and B lymphocytes, are altered compared to preflight levels. Lymphocytes from astronauts on board Soyuz 6, 7, and 8, Skylab 2, 3, and 4, and Salyut 4 exhibited poor response to mitogenic factors (25), the substances that induce the immune response. Therefore, the cells, experience a reduced functional capacity in microgravity conditions.

The immune system, like other body systems, responds dynamically to varying conditions, which may explain why results from one study contradict those of others. Studies conducted as early as the Skylab program suggest that impaired immune function during spaceflight is closely linked to the endocrine system and is particularly affected by corticosteroids and catecholamines (25); generally, this implies that changes in other regulatory mechanisms could closely affect immune function. In addition, results from *in vitro* studies may not parallel results from *in vivo* studies, indicating that the physiological environment plays an integral role in maintaining immunologic integrity (26).

It has been shown in several studies that stress has a considerable influence on immunity. Astronauts experience psychological and physical stresses that may result in reactivating latent viruses during space flight, potentially increasing the risk of infection among the crew. A study done on the amount of Epstein-Barr shedding pre-, in- and postflight on the Shuttle and Mir crews, showed that the virus, normally latent in most humans, was higher in samples taken before launch (27). Although results from this study suggest that stress levels are higher before than during or after flight, reactivation of latent viruses, combined with depressed immune response in flight, poses a threat to both short- and long-term missions in the event of injury or infections.

Because immune function and activity are interdependent on other systems of the body, there are implications for exploring countermeasures to mitigate the changes and for studying how pharmacological substances interact with the immune system in microgravity (18). Similar changes in immune response were reported in wintering personnel in Antarctica and, it is thought, are due to confinement and isolation. Carefully planned studies will be required to shed further light on this important issue, which could affect the development of chronic and debilitating diseases.

Hematologic, Fluid, and Electrolyte Changes. The headward shift of fluids in weightlessness and the resulting decrease in circulating blood volume are responsible for many of the physiological changes that occur during adaptation to spaceflight conditions. As has been discussed, they directly affect the functioning of the cardiovascular system. They also have several effects on the composition of body fluids, especially blood (Table 1). The most significant hematologic changes involve a reduction in plasma volume, alterations in red blood cell (RBC) mass,

and changes in the distribution of RBC shapes. From the time of the early Gemini and Vostok missions, a postflight decrease in total RBC mass has been observed in nearly all U.S. and Soviet crew members. There is a gradual decrease, losses average about 9% of the total RBC pool during the first 30 to 60 days in flight, and values range from 2–21%. Cosmonauts who participated in missions of 18 days to 6 months have shown a postflight decrease in erythrocyte counts that returned to baseline values within 6 weeks (22).

The magnitude of the RBC loss does not appear to relate to the length of missions longer than 14 days. Changes in RBC are also accompanied by changes in the shapes of erythrocytes, although these alterations do not seem to affect crew health or function in flight and are rapidly reversed postflight. The weight of evidence now suggests that the loss of RBC mass is due, instead, to insufficient circulating erythropoietin in combination with neocytolysis, or a decreased survival rate of newly formed RBCs (7,15). The decrease in RBC mass is effectively masked by a simultaneous, rapid decline in plasma volume (4–16% from preflight values) such that the ratio of cells to plasma remains roughly normal (18).

The microgravity-induced fluid shift produces at least a transient increase in central blood volume (18). Research from ground-based bed-rest studies suggests that the stretch receptors in the left atrium interpret this as an increase in total circulating blood volume and trigger a compensatory loss of water, sodium, and potassium from renal tubules. This is the first event in a series of fluid and electrolyte shifts that occur during the adaptation to weightlessness. So far, early diuresis has been observed only in bed-rest studies. It is difficult to demonstrate during spaceflight because of the problems involved in accurately documenting urine volumes early in flight while water intake is usually reduced due to SMS. Additional findings include in-flight increases in the urinary output of sodium, potassium, and chloride, an in-flight decrease in antidiuretic hormone, and reduced postflight excretion of sodium. Fluid retention has also been a consistent finding in cosmonauts after Soyuz flights, but it was found that excretion of potassium and calcium increases. Alterations in electrolytes are believed responsible for cardiac arrhythmias on Apollo 15 and subsequent U.S. and Soviet missions (7).

Psychological Health. The spaceflight environment consists of many elements that, even if experienced separately, are both physically and mentally challenging. A confined living space, high public interest and visibility, isolation from family and friends, crowded or often unappealing spacecraft conditions, and requirement for strong group dynamics are but a few of the obstacles that cosmonauts and astronauts face. These factors are compounded by the physical fact that light/dark cycles are altered in orbit, which affects circadian rhythm and is evidenced by disrupted or insufficient sleep (28). The most prevalent psychological events reported by crew members include high levels of stress or tension, anxiety often demonstrated as annoyance at other crew members or ground support personnel, decreased levels of concentration, emotional instability including mood elevation or depression, and general fatigue.

As increasingly longer missions become feasible, psychological and behavioral support has become an element of both the American and Russian space medicine programs. Often, support takes the form of comparatively small changes

in operations or scheduling that minimize crew requirements and permit crew members some flexibility in arranging their work/rest cycles. For example, astronauts have reported considerable improvements in outlook and performance when communications with family members or friends are provided regularly. Careful planning for work/rest cycles, proper recreation, nutrition, and interpersonal and family communications are essential in maintaining psychological health during long-duration missions. The Soviet/Russian program has developed an elaborate system for its crews, which is now adapted to the ISS program (7).

Time Course of Adaptations

The human body is exquisitely sensitive to changes in its surroundings and reacts to such changes with equal precision. Modest changes in gravitational force, for example, as a sitting person stands or a sleeping person awakens induce a host of regulatory or adaptive mechanisms to ensure that blood consistently reaches all extremities. A more significant change to the gravitational environment—such as the microgravity of spaceflight—clearly challenges the body's homeostasis to a much greater extent (6).

The earliest orbital flights were conducted in small capsules and lasted only a few hours or days. Within these first human-rated spacecrafts, the limited capacity for movement and the short exposure to microgravity meant that crew members mainly reported rapid onset of adaptation (1). As mission duration increased well beyond several days into months and even years, crews are now faced with further adaptive events and new physiological challenges; adaptation to spaceflight is neither instantaneous nor consistent, but instead depends on individuals, mission duration, and operational activities (6). Despite these differences, all crew members who return from both short- and long-duration flights report two periods of adaptation that occur after the transition from one gravitational environment to another.

The first is experienced upon launch and entry into orbit. Some symptoms manifested early in the mission abate as adaptation is resolved. The sensory conflict produced by the visual and vestibular systems is one example that is limited to the first 3 to 8 days of a mission.

The return to Earth's gravity requires a second period of adaptation, which again presents a significant challenge to crew activity and safety. Orthostatic intolerance stems from the cardiovascular deconditioning and cephalic fluid shift that occurs in response to microgravity; many crew members report presyncopal or syncopal episodes, that is, dizziness or fainting, upon return to 1g. Neuromuscular and neurovestibular adaptations produce postflight disequilibrium (including marked vertigo in some cases) and gait disturbances; both clearly limit coordinated maneuvers and interfere with nominal or contingency egress (13,21). Cosmonauts from long-duration Russian missions of 8 months have required more than 4 weeks of rehabilitation to function normally (29). Physical performance also declines as a result of significant and sustained loss of bone and muscle mass, documented at 10–20% of preflight levels during extended-duration missions.

Bone and connective tissue changes, for example, begin as early as 1 week into a mission and can continue for more than a year. These changes are not typically apparent in flight, but are instead demonstrated upon return to Earth as locomotor problems, bone frailty, and increased risk of kidney stones (22).

Challenges for Exploration-Class Missions

As human spaceflight continues beyond low Earth orbit, health monitoring and health maintenance through appropriate countermeasures will become more discrete and seamless in the spacecraft of the future. Crewmembers may well monitor their own medical status, evaluate environmental health, assess risks, and then direct automatic correction or restoration of an anomaly. The opportunity for novel or previously unexplored countermeasure approaches, including artificial gravity, could well alter what are currently considered the most dire biomedical challenges of human spaceflight. The crew of the International Space Station and future spacefarers will be just as dependent as their forebears on a thorough understanding and mitigation of these challenges (30).

BIBLIOGRAPHY

1. Swenson, L., J. Grimwood, and C. Alexander. This New Ocean: A History of Project Mercury. NASA SP-4201, U.S. Government Printing Office, Washington, DC, 1989.
2. Mark, H. *International Space Science, Physics in a Technological World*. American Institute of Physics, New York, 1988.
3. Froelich, W. Apollo-Soyuz, NASA EP-109, U.S. Government Printing Office, Washington, DC, 1976.
4. Souza, K., R. Hogan, and R. Ballard (eds). *Life Into Space 1965–1990*; Souza, K., G. Etheridge, and P.X. Callahan, 1991–1998 NASA/RP-1372 and NASA/SP-2000-534. NASA Ames Research Center, Moffett Field, California, also available at <http://lifesci.arc.nasa.gov/lis/index.html>.
5. Fazio, G. Vacuum, temperature, and microgravity. In S.E. Churchill (ed.), *Fundamentals of Space Life Sciences*, Krieger, Malabar, FL, 1997.
6. Nicogossian, A., C. Leach Huntoon, and S. Pool (eds). *Space Physiology and Medicine*, 3rd ed. Lea & Fibiger, Philadelphia, 1994.
7. Nicogossian, A., O.G. Gazenko, A.I. Grigoriev, and S.R. Mohler (eds). *Space Biology and Medicine*, Vol. I–III. AIAA and Nauka Press, 1997.
8. Badhwar, G. The radiation environment in low-Earth orbit. *Radiat. Res.* 148: S3–S10 (1997).
9. Straume, T., and M. Bender. Issues in cytogenic biologic dosimetry: Emphasis on radiation environments in space. *Radiat. Res.* 148: S60–S70 (1997).
10. The United States Prepares for the International Space Station. The Phase I Program 1998 <http://www.hq.nasa.gov/office/olmsa/>.
11. Eckart, P. *Spaceflight Life Support and Biospherics*. Microcosm Press and Kluwer, Dordrecht, 1996.
12. Manzey, D., B. Lorenz, and V. Poljakov. Mental performance in extreme environments: Results from a performance monitoring study during a 438-day spaceflight. *Ergonomics* 41 (4): 537–559 (1998).

13. Greenisen, M., J. Hayes, S. Siconolfi, and A. Moore. Functional performance evaluation, C. Sawin, G. Taylor, and E. Smith (eds), Extended Duration Orbiter Medical Project Final Report 1989–1995. NASA SP 1999-534. National Aeronautics and Space Administration, Houston, TX, 1999.
14. Goldberger, A., M. Bungo, R. Baevsky, B. Bennett, D. Rigney, J. Mietus, G. Nikulina, and J. Charles. Heart rate dynamics during long-term space flight: Report on Mir cosmonauts. *Am. Heart J.* 128 (1): 202–204 (1994).
15. Alfrey, C., L. Rice, M. Udden, and T. Driscoll. Neocytolysis: Physiological down-regulator of red-cell mass. *Lancet* 10; 349 (9062): 1389–1390 (1997).
16. Charles, J., J. Fritsch-Yelle, P. Whitson, M. Wood, T. Brown, and G. Fortner. Cardiovascular deconditioning, C. Sawin, G. Taylor, and E. Smith (eds), Extended Duration Orbiter Medical Project Final Report 1989–1995. NASA SP 1999-534. National Aeronautics and Space Administration, Houston, TX, 1999.
17. Wolthusius, R., S. Bergman, and A. Nicogossian. Physiological effects of locally applied reduced pressure in man. *Physiol. Rev.* 54 (3): 566–595 (1974).
18. Leach-Huntoon, C.S., A.I. Grigoriev, Y.V. Natochin. *Fluid and Electrolyte Regulation in Spaceflight*, Science & Technology Series, Vol. 94. AAS, Univelt Inc., San Diego, California, 1998.
19. Calvin, M., and O.G. Gzenko. *Foundations of Space Biology and Medicine*. NASA U.S. Government Printing Office and Nauka Acad. Press USSR, 1975.
20. Graybiel, A., E. Miller, J. Homick. Equipment M131. Human vestibular function, R. Johnston, and L. Dietlein (eds). Biomedical Results From Skylab. NASA SP-377. U.S. Government Printing Office, Washington, DC, 1977.
21. Guedry, F., A. Rupert, and M. Reschke. Motion sickness and development of synergy within the spatial orientation system. A hypothetical unifying concept. *Brain Res. Bull.* 15; 47 (5): 475–480 (1998).
22. Sawin, C. Biomedical investigations conducted in support of the Extended Duration Orbiter Medical Project. *Aviation Space Environ. Med.* 70 (2): 169–180 (1999).
23. Whittle, M.S., R. Herron, and J. Cuzzi. Biostereometric analysis of body form. R. Johnston, and L. Dietlein (eds), Biomedical Results from Skylab. NASA SP-377. U.S. Government Printing Office, Washington, DC, 1977.
24. Nicogossian, A., S. Pool, and C. Sawin. Status and efficacy of countermeasures to physiological deconditioning from space flight. *Acta Astronautica* 36 (7): 393–398 (1995).
25. Kimzey, S. Hematology and immunology studies, R. Johnston, and L. Dietlein, (eds), Biomedical Results From Skylab. NASA SP-377. U.S. Government Printing Office, Washington, DC, 1977.
26. Sherr, D., and G. Sonnenfeld. Response of the immune system to spaceflight. In S.E. Churchill (ed.), *Fundamentals of Space Life Sciences*. Krieger, Malabar, FL, 1997.
27. Payne, D., S. Mehta, S. Tying, R. Stowe, and D. Pierson. Incidence of Epstein–Barr virus in astronaut saliva during spaceflight. *Aviation Space Environ. Med.* 70 (12): 1211–1213 (1999).
28. Manzey, D. and B. Lorenz. Mental performance during short-term and long-term spaceflight. *Brain Res. Brain Res. Rev.* 28 (1–2): 215–221 (1998).
29. Vico, L., P. Collet, A. Guignandon, M. Lafage-Proust, T. Thomas, M. Rehaillia, and C. Alexandre. Effects of long-term microgravity exposure on cancellous and cortical weight-bearing bones of cosmonauts. *Lancet* 355 (9215): 1607–1611 (2000).
30. Nicogossian, A. and D. Pober. The future of space medicine. *Acta Astronautica* 49 (3–10): 529–535 (2001).

ARNAULD NICOGOSSIAN
NASA Headquarters,
Washington, DC

BIOMEDICAL SUPPORT OF PILOTED SPACEFLIGHT

Recent progress in the conquest of space and in piloted cosmonautics is the result of developments in space technology and hardware and, to a significant extent, has also depended on the solution of complex biomedical problems and other achievements in space biology and space medicine. During the period under discussion, the duration of space flight has increased to 14.5 months for men and 6 months for women. The long period during which the Mir orbital station was used (15 years) has generated unique experience in solving biomedical problems to ensure the safety and efficiency of increasingly long spaceflights.

Space medicine is basically a type of prophylactic medicine. Its duties include predicting all of the physiological consequences of exposure to spaceflight factors and preventing or curtailing likely disruptions in the functioning of various systems.

Support of piloted spaceflights involves a complex set of biomedical, technical, and organizational measures directed at creating the conditions required if normal human vital processes are to occur in space and using special means and methods to maintain cosmonaut health and performance at the level needed to ensure completing the standard flight program and for high crew efficiency if contingency or emergency situations occur.

Many years of research conducted by the Institute of Biomedical Problems jointly with a number of other organizations (see section on Space Life Sciences) has resulted in creating an efficient biomedical piloted spaceflight support system. This system includes

- cosmonaut medical selection and training;
- medical monitoring of cosmonaut status during spaceflight;
- providing medical care on board orbital stations;
- postflight cosmonaut rehabilitation;
- preventing the adverse effects of weightlessness; developing ways to prevent cardiovascular and musculoskeletal deconditioning, disruptions of fluid-electrolyte metabolism, and sensory disorders;
- providing cosmonauts with food and water during flight;
- life support for spacecraft;
- maintaining radiation safety;
- maintaining cosmonaut safety during EVAs;
- sanitary and hygienic support of cosmonauts;
- developing optimal work-rest schedules and systems for psychologically supporting crews.

The overall program of measures performed by personnel working in space medicine is very extensive and includes three phases: preflight, in-flight and postflight (see Table 1).

Table 1. Program of Work in Space Medicine

Preflight measures	In-flight measures	Postflight measures
Identification of latent diseases and insufficiencies in compensatory physiological mechanisms during space-flight selection and training; preflight preventive treatment, if necessary quarantine or observation and other antiepidemiological measures; preventive surgical interventions; measurement of individual sensitivity to drugs	Prophylactic, diagnostic, and therapeutic procedures onboard the spacecraft; performance of medical procedures directed at physiologically preparing cosmonauts for the powered stages of flight (insertion into orbit and touch- or splash-down); if necessary, arranging emergency return to Earth	Medical monitoring and, if necessary, provision of medical care to cosmonauts after landing; designing and implementing rehabilitation measures; medical observation and development of measures to aid rapid crew-member postflight adaptation to conditions on Earth; medical examination between flights and at retirement to identify remote consequences of space flight

Major Components of the Biomedical Support of Piloted Spaceflight

Medical Selection and Cosmonaut Training. Clinical physiological examinations of cosmonauts are undertaken to enable expert evaluation of their health in various phases of flight preparation and during the postflight period. The number of examinations and the intervals between them are specified by a special program (1). Before being accepted in the training group, candidates for crew membership undergo an inpatient examination, whose results are used to determine whether or not their health allows accepting them for the crew-flight-training program.

During the training phase, cosmonauts undergo dynamic monitoring, allowing specialists to

- detect latent incipient stages of disease;
- study personality characteristics and assess an individual's functional and potential physiological capacities;
- develop recommendations determining the sequence and time line for various types of training and also measures for increasing physiological tolerance to exposure to spaceflight factors (2).

During the training period, crews periodically undergo detailed clinical physiological examinations to

- identify the characteristics of each cosmonaut's physiological and psychological reactions to specific types of training;

- on the basis of health status and psychological traits, determine cosmonauts' fitness for spaceflights of varying duration.

The medical commission uses the results of this comprehensive examination to decide whether to qualify crews for spaceflight.

Medical Monitoring of Cosmonaut Status During Spaceflight. Medical monitoring is an important component of the set of measures used to ensure crew safety during spaceflight. Such monitoring makes it possible to identify, analyze, and evaluate functional physiological changes exhibited by a cosmonaut, to assess whether use of prophylactic measures is indicated, and to select the optimum schedules for using them. An unusual feature of in-flight medical monitoring is that the patients are healthy, physically fit individuals. In addition, unlike medical monitoring on the ground, in-flight monitoring is remote—because the subject and the medical personnel are separated by vast distances. Under these conditions, sources of information consist of data on the status of a cosmonaut's physiological systems, the microclimatic parameters in the spacecraft or suit, the contents of radio communications between the crew and the Flight Control Center, and also special radio communication by the cosmonauts concerning how they feel during the flight. Telemedicine sessions, observations of crew actions, and monitoring how well they are performing their programmatic tasks serve as additional sources of valuable information (3).

The goals of in-flight medical monitoring include ongoing real-time monitoring of cosmonaut health status; identifying functional changes in physiological systems, and also of pathologies arising in flight; predicting potential for further continuance of the flight; planning and managing medical tests and prophylactic measures that facilitate retention of performance capacity in-flight and postflight; as well as monitoring the cabin environment, radiation conditions on the flight path, and adherence to the work-rest schedule, all of which are described in the respective portions of this article (4). In addition to this operational monitoring, periodically, crew members are subjected to more in-depth medical examinations, including administration of provocative (loading) tests to assess levels of functional physiological reserve capacity. Furthermore, the medical monitoring system, if necessary, can be used to perform an immediate emergency medical examination and can be shifted to a mode of continuous recording of physiological parameters.

Means and Methods of Medical Care in Flight. The means and methods for providing medical care if cosmonauts develop illnesses or health disorders are considered of prime importance within the system for operational medical support of piloted flights.

The spaceflight medical support system is tasked with preventing disease, injury, and exposure to toxic substances and penetrating radiation, and also various different functional disorders evoked by exposure to spaceflight factors. The medical care system has the goal of providing timely diagnosis of health disorders and effective aid to any crew member who requires treatment during preflight training, in flight, or after return to Earth (5).

Although only healthy individuals are permitted to fly and all sorts of prophylactic measures are employed in flight to minimize the negative physiological effects of spaceflight factors, it is still not possible to preclude completely the

occurrence of disease or other conditions that require medical treatment. Any such case is analyzed to identify possible shortcomings in the measures prescribed by the medical flight support system, and then the appropriate recommendations are derived.

The stressful operator activity, the metabolic shifts that occur in weightlessness and the changes in the reactivity of physiological systems may decrease resistance to adverse factors and damaging effects. This may lead to the development of neurasthenic symptoms (decrease in work capacity, susceptibility to fatigue, irritability, sleep disturbances), various inflammatory diseases and allergic states, and metabolic disorders of organs and tissues. Nor are such diseases as acute appendicitis or inflammation of the gall bladder or pancreas precluded during space flight because their occurrence is extremely difficult to predict even under normal conditions on Earth. Finally, emergency situations on board a spacecraft may lead to injuries, barotrauma, or poisoning.

The list of states that requires using prophylactic and/or therapeutic measures during spaceflight must include "motion sickness," the neurasthenic syndrome, local inflammatory diseases, and minor traumas.

The isolated nature of their living conditions, the unique conditions of their environment, and the limited number of crew members compel the cosmonauts themselves to bear responsibility for diagnosing illness and providing medical care.

All cosmonauts are adequately trained to provide medical care. If necessary, crew members, using onboard documentation, the computer database, and the capabilities of the telemedicine system can employ the appropriate onboard medical kit or medical procedures. Of course, the presence of a physician in the crew significantly facilitates medical monitoring and medical care and enhances their efficacy.

A special mobile unit equipped with all of the necessary medical instruments, which can be deployed at the reentry vehicle landing site, has been developed for use if it is necessary to give cosmonauts medical care immediately after landing.

System to Prevent Adverse Effects of Weightlessness. The system for preventing the adverse physiological effects of weightlessness includes a set of medical and technical measures that supports long-term human residence and work in space, including fostering appropriate adaptation to flight conditions and readaptation after return to Earth while fully maintaining health and performance capacity.

Weightlessness, which induces a number of specific adaptive changes in various physiological systems, may lead to the development of functional changes and to some structural changes as well.

In essence, the prophylactic countermeasures used today (see Table 2) are directed at preventing or substantially attenuating these adaptive processes, primarily to safeguard and facilitate the process of readaptation upon return to normal gravity. The two main goals of the system of prophylactic measures are to make up for the deficit in motor activity and to compensate for the effects of fluid (blood and interstitial fluid) redistribution typical of weightlessness (6).

On long-term flights, cosmonauts follow a schedule of prophylactic measures that has been specially developed for that flight on the basis of duration,

Table 2. Counter Measures in Space Medicine

Class of methods	Mechanisms underlying effects	Procedures
Compensation for effects of blood redistribution in weightlessness	Induction of lower body blood pooling; stimulation of neuroreflex mechanisms regulating circulation in erect posture on Earth	Lower body negative pressure; occlusion cuffs
Physical exercise	Maintenance of conditioning in most important physiological systems; activation of venous pulsations, circulation facilitation from muscle contraction; activation of weight sensors and proprioceptors; maintenance of motor skills needed to maintain vertical posture and perform locomotion after return to Earth	Physical exercise; (constant) loading suits; electric stimulation of muscles
Enhancement of orthostatic and acceleration tolerance	Maintaining hydration Preventing blood pooling in the lower body	Fluid-electrolyte supplements Anti-g suits
Alimentary correction	Correction of nutrients ingested	Dietary supplements containing minerals, amino acids, and vitamins
Pharmaceutical correction	Targeted effects on certain physiological functions to correct or prevent pathological symptoms	Numerous groups of drugs
Nonspecific prophylaxis	Methods fostering increased general physiological resistance to adverse effects	Endurance training; psychosomatic regulation of functions; breathing of gas mixtures with diminished oxygen
Correction of adverse external factors	Decrease in excessive stress on physiological systems and functioning during spaceflight	Optimization of living environment, and work-rest schedule; means and methods of psychological support

flight phases, and the functional status of physiological systems. The following measures are included:

1. Constant wearing of the "penguin" constant loading suit during all waking hours in the intervals between physical exercise sessions, including during the performance of professional duties. This suit applies lengthwise axial loading to the musculoskeletal system (from the shoulder girdle to the foot) to simulate weight loading and reproduces some degree of deformation and stimulation of the resistance and muscle receptors.

2. Two hours daily of physical exercise on the inboard exercise machines (VB-3 bicycle ergometer and UKTF treadmill) to maintain conditioning in the most important systems and retain overall physiological work capacity, activate venous pulsation and the circulation facilitating effects of muscle contractions, stimulate mechanoreceptors, and also foster retention of skills needed for maintaining a vertical posture and for locomotion. Physical exercise generally involves the UKTF apparatus, which consists of a treadmill equipped with a system of straps, individually fitted loading suits, special shoes, and an elastic harness. The treadmill permits cosmonauts to walk, run, jump, do knee bends, and lift “weights,” thus reproducing constant static loading along the vertical axis of the body and the effect of maintaining upright posture in gravity.
3. Electric stimulation of muscles—high frequency (using the Tonus-3 stimulator) and low-frequency (using the Myostim)—preferably after physical exercise, to maintain muscle strength, and static and dynamic endurance.
4. Use (in accordance with a specially developed schedule) of the “Chibis” suit, which creates lower body negative pressure and thus in weightlessness, reproduces the hydrostatic blood pressure pattern characteristic of Earth. During the early period of adaptation to weightlessness (week 1 of flight), if they so desire, cosmonauts use “Bracelet” occlusion cuffs and the “Karkas” and “Kentavr” devices to decrease the severity of blood redistribution effects.
5. During the flight, fluid and salt supplements are ingested to facilitate fluid retention and prevent symptoms of dehydration, and thus, help to increase endurance to acceleration and orthostatic tolerance. During descent, special couches and anti-g suits that increase orthostatic tolerance and endurance of gravitational loading are used.
6. A balanced diet (containing salts, amino acids, and vitamin additives) is provided to combat possible deficits in the food.
7. Use of drugs to affect certain physiological functions to eliminate adverse symptoms or exert a corrective effect (preventing and/or treating “motion sickness,” maintaining of orthostatic tolerance, counteracting bone loss, and normalizing myocardial metabolism).
8. A set of psychological support measures to combat the adverse effects of spaceflight—isolation, a sealed, technogenic environment, excessive stress, and compulsory adherence to a nonoptimal work-rest schedule (7).

Preflight, to increase general physiological tolerance of adverse effects, nonspecific prophylactic measures are used, including endurance training, physical and special training, sleeping in head-down position, and breathing gas mixtures with diminished oxygen content.

Prophylactic measures directed at preventing or partially compensating for undesirable changes caused by weightlessness play an important role in maintaining cosmonaut performance level in flight and in ensuring their safe return to Earth.

Life Support on Board Spacecraft. A life support system (LSS) is a set of devices and systems, as well as supplies of food and other substances, required to maintain vital processes (metabolism) and human performance in pressurized spacecraft cabins. The LSS maintains the atmosphere in the closed cabin at a preset chemical composition and physical parameters (pressure, temperature, humidity, rate of movement), satisfies the crew's need for food and water, and disposes of the wastes of the crew and other biological subjects. In accordance with these functions, the LSS is divided into a number of subsystems (components): air regeneration, water supply, food supply, thermal regulation, and sanitary and hygienic support. This is the structure of the typical LSS in the narrow meaning of the term. On long-term spaceflights, measures to maintain crew health and performance capacity are significantly enhanced. For this reason, the LSS also encompasses all devices and objects that satisfy the day-to-day cultural and aesthetic needs of the crew, provide physical exercise (weightlessness and the limited size of the inhabited portions of the spacecraft lead to a deficit in physical activity), and also provide radiation shielding.

LSS parameters are determined by the need to satisfy human requirements for food, water, oxygen, and for waste disposal. When LSSs are designed, all relevant factors are considered (purpose, type of spacecraft, duration of functioning, size of crew, characteristics of flight path, mass-energy constraints, safety, reliability, cost, and performance characteristics). The LSSs developed for Russian spacecraft are highly reliable, display stable performance under the influence of spaceflight factors, even in emergency situations, and are distinguished by their minimal power use, mass and size, and good maintainability.

As a function of the way they replace consumables, LSS subsystems are classified as either nonregenerative (based on stored supplies of the needed substances or supplies brought by transport spacecraft) or regenerative (the substances needed to support human life are recovered from biological wastes of humans and other spacecraft inhabitants) or mixed.

The LSS for the first spaceflight was designed on the "nonregenerative" principle and provided the cosmonauts with a store of equipment and onboard supplies whose nature and amount was based on the caloric needs and the nutritive balance required by a healthy human on Earth. But, as flight duration increased and information accumulated on the dynamics of human metabolism under spaceflight conditions, there was a growing tendency to design and implement integrated regenerative systems, which utilized wastes and by-products of the functioning of various LSS components. Creating a completely closed LSS, simulating the processes of Earth's biosphere, based on methods of abiogenic synthesis, remains a difficult challenge to this day. Partially closed LSSs, which regenerated water and the most important components of the atmosphere and also disposed of solid wastes, were developed for the flights of Salyut and Mir.

The formation of the spacecraft cabin atmosphere during flight is directly associated with the problem of atmospheric pollution. Sources of pollution may include construction materials and technological processes, as well as human waste products. The study of the biological effect of spacecraft atmospheric pollution is one of the more important problems that requires physiological and hygienic research. The practical result of such research is establishing of threshold limit values (maximal permissible concentrations) for a wide range of

polluting (toxic) substances and developing of techniques for removing them from the spacecraft's atmosphere (8).

So-called biological life support systems (BLSS) are based on a qualitatively new principle of environmental formation. The functioning of a closed biological system is based on the principle of repeated use of a relatively small initial quantity of chemical elements as the closed cycle transforms the substances of the system itself. To support human life, a BLSS must transform human waste products into food, oxygen, and water, that is, regenerate them. That such a system can exist in principle is obvious because the elements used by an adult organism are eventually emitted into the environment in the same quantities. Thus, for example, oxygen that is consumed oxidizing organic substances in food is emitted as a component of water and carbon dioxide, along with the hydrogen and carbon of the oxidized nutrients (9).

Systems based on biogenic synthesis are closest to conditions on Earth and thus to the biological needs of human beings, and are capable of self-regulation at all levels of the system by mutual correction of processes.

The first practical results in this direction were obtained in experiments on board Space Station Mir, when three complete developmental cycles of wheat "from seed to seed" were completed in the station's Svet greenhouse.

Cosmonaut Nutrition and Water Supply In-flight. Sufficient and well-balanced nutrition of crews is one of the most important ways to maintain health and high psychophysiological energy and is the source of positive emotions in spaceflight. As flight duration and the degree of crew isolation increase, job pressure and responsibility are enhanced, and stress situations become more frequent, the importance of a diet adequate for cosmonauts' physiological needs increases as well.

Onboard cosmonaut nutrition systems on Salyut and Mir were highly developed technological systems, including, aside from the food itself, the appropriate "infrastructure" to ensure reliable storage and enable the crew members to prepare and eat their meals under spaceflight conditions. The nutritional systems components were functionally linked to the other life support subsystems, especially the water supply system, waste collection and disposal system, the power supply system, and the system maintaining microclimatic parameters (10).

Food rations (daily or for the whole flight) were appropriate to the cosmonauts' caloric needs and contained the optimal amounts of nutrients (proteins, fats, carbohydrates, vitamins, and minerals). This was particularly true of the essential nutrients (certain amino acids, unsaturated fatty acids, and vitamins), which are insufficiently synthesized by the human body or not synthesized at all.

The major task of the water supply system was to ensure that the cosmonauts received a regular supply of water in quantities appropriate to their physiological needs to prevent the development of a fluid deficit in flight. Water is the largest component of the human body by weight, and its daily consumption exceeds the total consumption of oxygen and nutrients. This means that every day the cosmonauts were given ad lib access to a significant quantity of water of acceptable taste and odor, free from toxic contaminants and, if necessary, enriched with minerals.

On relatively short flights (up to 30 days), water supply systems based on supplies of potable water brought from Earth were preferred. As flight duration

increased, it became necessary to have an entire water generation cycle take place on board the spacecraft. For this purpose, the space stations were provided with systems for regenerating water from low concentrated (concentrates of atmospheric moisture and process water) and highly concentrated (human urine and wash water) solutions for repeated use. The problem of in-flight water regeneration also encompasses problems of conditioning the taste and the chemical, as well as the bacteriological, properties of the regenerated water before it is used for drinking or reconstituting dehydrated food (11).

Sanitary and Hygienic Support of Cosmonauts. Sanitary and hygienic support of cosmonauts includes a wide range of problems relating to maintaining conditions in the spacecraft cabin where they live and work that are conducive to crew comfort and well-being. This includes developing hygienic standards and prophylactic measures for maintaining the spacecraft crew's health and performance capacity, providing cosmonauts with clothing and means of personal hygiene, waste disposal, maintaining cosmonaut microflora in an optimal state, and maintaining the spacecraft atmosphere and surfaces (12).

The significance of the personal hygiene component of the biomedical flight support system has grown as flight duration has increased. Various technological processes and measures were used to satisfy crew-members' sanitary and hygienic needs. Here, the significance of hygienic procedures was dictated by hygienic and physiological considerations and also mainly by psychological/aesthetic, epidemiological, and possibly toxicological concerns.

The criteria for selecting personal hygiene devices and techniques included inducing a sensation of bodily cleanliness after the procedure had been completed and also the feeling of "refreshment" and psychological comfort.

Personal hygiene measures included four basic types of procedures:

- complete cleansing of the body;
- washing of certain portions of the skin;
- oral hygiene;
- haircuts, shaving, and nail care.

Dry and wet (moistened with special washing and cleaning solutions) wipes and towels were used to cleanse the body. The wipes provided adequate skin cleaning and refreshment. In addition, they could be used to wipe down the surface of the spacecraft cabin.

Oral hygiene was considered an important personal hygienic measure. It involved regular cleaning of the teeth and rinsing of the mouth. Various types of toothbrush, toothpaste and powder, toothpicks, mouthwash, and rinses were used for this purpose.

Hygienic treatment of the hair consisted of periodic haircuts and shaving of the beard and mustache. The main problem in hair care during space flight is preventing fragments of hair or beard from getting into the cabin atmosphere.

The saying that the best clothing is that whose presence cannot be felt is highly applicable to cosmonaut clothing, which must be comfortable for working and relaxation and should not impede or limit motion. The cosmonaut wardrobe consists of underwear, a flight suit, which the cosmonaut wears inside the

spacecraft, and a thermal suit. The fabrics specially selected for cosmonaut underwear were light and elastic, did not impede heat convection and radiation or evaporation of moisture from the body surface, and at the same time were strong enough to be worn for long periods of time and to allow attaching sensors for recording biotelemetric information.

The flight suit was fully compatible with the underwear and thermal suit. Its design allowed freedom of motion and made using sanitary facilities convenient. It was easy to put on and take off. One of the main functions of the flight suit was to maintain the cosmonaut's thermal balance by preventing both excess heat loss and accumulation of excess heat. The thermal suit is designed to be used on landing in a deserted spot under adverse climatic conditions. In addition, it may be used if the spacecraft air conditioning system does not function properly. The thermal suit set included shoes, which had to be light, strong, and to have good thermal insulation properties. This was achieved by selecting the appropriate materials and design, which also account for the fact that the cosmonaut would have been adapted to weightlessness (13).

In closed pressurized cabins of limited size, the presence of even small concentrations of harmful substances in the atmosphere may have a serious effect on human health and performance capacity. The sanitary hygienic subsystem included studies of sensory parameters and chemical analytic tests of the air, which made it possible to identify the nature and rate of the offgassing of harmful substances from polymers and wastes. One of the main sources of atmospheric pollution in a pressurized environment is the human body, which releases a large quantity of metabolic products through the lungs, skin, kidneys, and intestinal tract. The amount of volatile chemicals emitted by a person varies within broad limits and depends on a number of factors: the nature and quality of the diet, metabolic status, and the nature and intensity of work performed.

Polymer materials are a second, no less important, source of atmospheric pollution in a pressurized cabin. Polymer synthesis involves using a number of auxiliary compounds that belong to a number of different classes and have an extremely broad spectrum of toxic effects.

The concept of "wastes" encompasses the set of products that form as a result of human vital activity and the operation of the equipment installed in the spacecraft cabin and not subsequently used. Regardless of chemical and bacterial composition or physical and other properties, wastes have one basic property in common. They are one of the sources that pollute living and working areas with undesirable or harmful substances and foster the development of microflora, some of whose species can cause illness in crew members or damage equipment (14). Correct structuring of waste containment and disposal is one of the major requisite conditions for maintaining normal vital processes and high performance capacity in spacecraft crew members.

However, the technical implementation of these operations under the unique conditions of spaceflight encounters a number of difficulties. The great majority of wastes are gaseous or liquid. Experience with space stations Salyut and Mir has shown that collection and transport of substances in this state present great technical difficulties in weightlessness.

Research on the problem of microbiological damage (biodegradation) to structural materials was initiated when Salyut-6 was in use and continued on all

subsequent Soviet and Russian orbital stations. It was established that spacecraft microclimatic parameters, i.e. the presence of specific chemical contaminants in humidity condensate and anthropogenic pollution (with human metabolic products) are the stimulating factors for growth of bacteria and mold on the materials of the cabin interior and equipment. More than a 100 species of microbes—bacteria and fungi—have been isolated from the surfaces of these materials during long-term spaceflights. Among these were species that present potential danger to human health, the so called pathogenic saprophytes, which can grow actively on artificial substrates, and also nonpathogenic bacteria and fungi that damage (destroy and degrade) various materials (metals and polymers) and thus cause failures and disrupt instrument and equipment operation.

Because, cosmonauts may show symptoms of dysbacteriosis or develop states of immunodeficiency from the effects of spaceflight factors and the processes of pathogen recirculation are intensified in a small pressurized cabin, the risk of spread of infection among crew members increases. All of this motivated the development of a microbial spacecraft safety system, which stipulated, in particular, that during preparations for flight, only structural materials that had been, in ground-based simulations, most resistant to microbial action were to be selected and that special disinfecting measures be undertaken. During spaceflight, the system called for treating spacecraft cabin surfaces, including the spaces behind instrument panels, with wipes moistened with special antibacterial and antifungal agents (fungistats). Experience with long-term flights on Salyut and Mir demonstrated that this system is highly effective.

Maintenance of Radiation Safety in Spaceflight. The radiation safety system is a set of means and measures directed at preventing and precluding the adverse effects of ionizing radiation (for example, during powerful solar flares or flights in Earth's radiation belt) on cosmonauts. These means and measures include physical screening of inhabited spacecraft modules, additional local screening of cosmonauts (radiation shelters), pharmacological and chemical protection of cosmonauts, inboard radiation monitoring devices, and the results of monitoring the radiation situation by the Solar Service (15).

Measures directed at ensuring crew safety include predicting the level of cosmonaut radiation exposure during the planned flight, developing initial recommendations for constructing piloted spacecraft, analyzing the radiation conditions in the flight path, radiation monitoring in the spacecraft cabin and station orbits, evaluating of levels of radiation exposure; developing recommendations to keep irradiation from exceeding the threshold limit dose, and providing cosmonauts with complete and current information.

The high biological interaction of various types of cosmic radiation makes them dangerous. For this reason, a research study was undertaken to determine acceptable levels of radiation exposure and to develop means and methods for prevention and for shielding cosmonauts from cosmic radiation.

By now, many techniques and devices have been developed to measure absorbed dose, dose equivalent, particle flow, linear energy transfer spectra, charge and energetic spectra and spectra of particle mass; each serves a strictly defined measurement function.

Radiation monitoring systems may be classified as active or passive. In active systems (including tissue equivalent ionization chambers, microdosimeters,

and particle spectrometers), instrument readings are recorded by crew members in orbit or are telemetered to Earth in realtime. In passive systems (including thermoluminescent dosimeters and dielectric track detectors), the readings are recorded and analyzed after the spaceflight has been completed.

Evaluation of radiation risk involves assessing the likelihood of specific adverse somatic effects (ASE) on human health as a result of exposure to ionizing radiation. The conception of risk from the effects of ionizing radiation assumes that the likelihood of developing ASE is directly proportionate to dose equivalent. In this case, the likelihood of ASE is the product of two contingent probabilities: the likelihood that a person will be exposed to a given dose equivalent and the likelihood that the dose equivalent will provoke ASE.

The risk to humans of cosmic radiation in flight may be minimized by a number of measures to decrease the likelihood of ASE to a justified (minimal reasonably acceptable) level, that is, according to the "ALARA" (as low as reasonably achievable) principle. The concept of a reasonably acceptable risk makes it possible to decrease it, comparing advantages and disadvantages, under the assumption that certain threshold limits on a momentary dose of irradiation are the upper limits of the safe level of exposure. It follows from this that a dose limit exists (the lowest level) that is absolutely unacceptable to exceed if there is to be any further exposure. Thus, it is not sufficient to decrease radiation exposure to a level below the dose limit; rather risk must be limited by reducing all radiation exposures to the minimally reasonably acceptable level.

In accordance with the standard dose limits adopted in Russia, the amount of irradiation to which a cosmonaut's hemopoietic organs have been exposed throughout the entire period of his career must not exceed 1 Sv (16). This dose limit has been established to limit the adverse remote effects of cosmic radiation. To avoid immediate radiation effects during flight, which may decrease in-flight performance capacity, dose standards for shorter periods have also been adopted. For example, the annual dose limit is 0.5 Sv, and the monthly dose limit is 0.25 Sv. On 1-year orbital flights, the total dose did not exceed 0.2–0.25 Sv, although there is some small probability that this value should be increased significantly to account for powerful solar proton events. Thus, the adopted dose limits and the radiation conditions characteristic of near-Earth orbit permit increasing space flight durations on these paths to up to 4–5 years.

In practice, radiation risk may be decreased by regulating the amount of time an individual spends in the cosmic radiation field and also by designating an area within the spacecraft that has a low radiation dose rate where the cosmonauts spend the majority of their time. Radiation risk may be decreased during the spacecraft design phase by selecting the best materials for radiation screening to prevent particles from penetrating the interior of the spacecraft cabin and also by minimizing induced radioactivity. The total whole body radiation dose as well as the dose to certain organs can be significantly decreased by local screening of these organs and areas of the body. Moreover, medical prevention and treatment methods are being developed (use of radioprotectors, prophylactic biomedical drugs, and postradiation therapy) that may enhance reparative processes in affected tissues. There is also a possibility that proactive genetic selection of cosmonauts could be used (selection of individuals who are

highly resistant to the effects of radiation or whose tissues can rapidly recover from radiation damage).

Psychological Support of Long-Term Flights. The increase in spaceflight duration and the complexity of flight missions have substantially increased the priority of the human factor in the “cosmonaut–spacecraft” system. Solution of problems relating to human psychological stability on long flights involves considering a large number of behavioral factors: psychological needs; subjective states; anxieties; interactions with colleagues on the crew and on the ground, for example, at the Flight Control Center; role-based relationships; work planning; criteria for success; and an external motivation system. The difficulty of maintaining psychological stability in a crew increases as flight duration begins to be measured in years.

One of the problems of space psychology is how to increase the psychological and professional reliability of cosmonauts. The solution of this problem will require improving the means and methods of selection, training and crew formation, and of evaluating cosmonauts’ psychological status. Prevention and correction of psychological disadaptation is extremely important here and must entail investigating the characteristics of group dynamics and the chronobiological aspects of adaptation, as well as optimizing cosmonaut professional performance (17).

Improvement of crew living and working conditions on board the spacecraft attenuates the psychogenic consequences of living in an artificial environment. During preflight training, a cosmonaut masters the necessary professional skills and knowledge and the skills involved in group dynamics. On this basis and also based on their own beliefs and expectations, cosmonauts construct their own individual representation of the upcoming spaceflight.

One particularly significant psychological problem that arises in connection with the increased heterogeneity of space crews in nationality, gender, profession, and other characteristics is the problem of optimizing the psychological climate in a heterogeneous crew, ensuring that effective independent group decisions are made, and that international space crews interact successfully with national flight control centers. In this context, the crew training phase, during which cosmonauts must develop mutual trust, a common system of values, and a strategy for managing and resolving problems as a team, takes on additional significance (18).

During the first 4–6 weeks of flight, cosmonauts come to terms with the living environment in the spacecraft. During this period, they must cope with new sensations, impressions, and characteristics of movement. In this situation, cosmonauts require additional time to prepare for and implement their professional tasks and to set up interactions with specialists in ground services. This may lead to time pressure and fear of not completing tasks on schedule, which creates conditions conducive to emotional stress and fatigue.

Next comes a phase of physiological and psychological stabilization; however, the crew members begin to feel an intense desire for new information as a result of the “sensory deprivation” they are experiencing. The novelty of spaceflight begins to lose its significance for them, as they get used to the unfamiliar living and working schedule and conditions. The effect of these factors can lead to

diminished psychological tonus, development of symptoms of debilitation, and disruption of the “sleep–waking” cycle.

In the period 15–30 days before landing, crew members enter another phase of psychological adaptation—emotional reorientation—as they begin to focus on imminent return to Earth.

The shift to long-term space flights, the way to which was paved by technological progress and the new capacities of space technology in the 1970s, required well-grounded safeguards that crew members would retain their health, high performance capacity, and the ability to work without “breakdown.”

As part of medical flight support, a great deal of attention is paid to monitoring the crews’ work–rest schedules. Information is collected about planned and actual cosmonaut schedules, these data are analyzed immediately, and scientifically justified suggestions and recommendations are generated for immediate adjustment of the schedule on one- and multiday time scales. The work–rest schedule is then revised, taking into account the crew’s status and progress in flight program implementation.

Maintenance of normal rhythms in physiological functions takes on critical significance on long-term flights. The planned work–rest schedules of Salyut and Mir cosmonauts were based on the familiar 24-hour schedule without displacing the “sleep–waking” cycle. However, in a number of instances, the need to perform such critical tasks as launch, docking, EVAs, and others compelled occasional shifts in the phases of the cycle.

Each shift of this kind represented stress and led to additional pressure on regulatory systems. In cases like this, special measures were taken to minimize the adverse consequences of these shifts and prevent desynchronization.

In addition to the duration of crew work shifts, each member’s interest in various types of task was studied, as was the relationship between their level of motivation and the quality of their work.

A number of effective quantitative methods were developed to aid in objective assessment of cosmonauts’ psychological status on long-term flight, and these have been used successfully in medical support of piloted flights, including flights of international crews.

Although the structuring of the crew’s lives on long-term space flights is limited by technical capacities, medical personnel and psychologists have labored intensively to improve conditions for living and working in space and to attenuate the psychogenic consequences experienced by crew members because of shortcomings in the artificial living environment.

The concept of “psychological support” was introduced to space medicine in connection with the support of space-station flights of increasing duration. Cosmonauts who have made long-term flights unanimously acknowledge that the standard system of psychological support is an important factor in maintaining a normal sense of well-being and performance capacity under such conditions. Generally, the aspect that has the most important role is the opportunity to have private conversations with their families and unstructured communications sessions with their relatives, friends, and various public figures.

The system of psychological flight support for cosmonauts is directed at optimizing the psychological status and performance capacity of healthy individuals, prevention of psychological and psychophysiological impairments, and

support of harmonious psychological interactions within the crew. Only through effective use of psychological and personal reserves of strength during a flight is it possible to avoid developing undesirable neurophysiological changes (diminished performance capacity, disruptions of sleep, debility, and conflict stress) that occur as a result of adverse psychological spaceflight factors: heightened risk, sensory deprivation, monotony, heightened responsibility for performance of flight operations, and the limited and imposed nature of social contacts.

Psychological support, which is generally not based on the use of drugs, makes it possible to compensate effectively for the deficit in social contacts (teleconferences with relatives and friends), the information deficit (news broadcasts and Internet access), and gives the cosmonauts the opportunity to feel the importance and interest those on the ground attach to the results of their work and their constant concern and attention to them as individuals (through packages containing their favorite books, films and music, and congratulations on dates of personal importance). Particular significance is attached to compensating for the loss of accustomed terrestrial stimulation—landscapes and sounds of nature—which are reproduced on special video and audio programs, produced with help from cosmonaut friends and family members. If necessary, cosmonauts can talk to psychologists on a confidential channel. Constant work with the families of crew members makes it possible to optimize the psychological climate surrounding each cosmonaut, to prepare family members for communicating with the space station, and to facilitate subsequent postflight psychological rehabilitation.

Space medicine and space psychology have developed objective methods for daily psychological monitoring of cosmonaut psychophysiological status and performance capacity, as well as of the psychological climate of the crew as a whole. Diagnoses are made by analyzing radio conversations with the crew as experts make ratings on specially developed scales. Significant diagnostic information regarding performance capacity is provided by the onboard psychodiagnostic system. In addition, each cosmonaut's pattern of psychological and emotional reactions to various situations, including extreme ones, has been charted preflight. The diagnostic information obtained is used to plan problem-oriented psychological support measures; predict changes in psychological status and work capacity in subsequent phases of the flight; and provide recommendations to flight directors on the advisability of a cosmonaut performing key operations—night work, docking, and EVA.

Methods of prevention and correction include the following:

- special preflight training;
- measures to optimize professional performance;
- formulation of positive feedback that acknowledges the success of tasks performed;
- measures to counteract the “asthenic syndrome” (dysthymia) and eliminate nonspecific anxiety components, including the use of drugs.

Support of Cosmonaut Safety During EVAs. EVAs are an important and effective operation performed during space flight. During the period Soviet

and Russian spacecraft were used between 1965 and 2000, a total of 96 EVAs were performed by 51 different cosmonauts (including one woman). Cosmonauts performed an enormous amount of work during these operations; they conducted unique scientific experiments, transported and mounted large structures on the space station exterior, performed various repair and debugging tasks, tested self-contained manned maneuvering units, and inspected depressurized modules. The maximum duration of an EVA from opening to closing of the access hatch was 7 hours 14 minutes by A. Solvyev and N. Balandin in 1990 (19).

The EVA medical support system includes monitoring during training, EVA implementation, and post-EVA. At present, there are no specific medical requirements for crew members who conduct EVAs. They undergo obligatory training on a technical trainer, learn about the EVA suit and study a number of operations that crew members will have to perform.

Four basic technical trainers are used: a setup enabling training in the absence of weight loading (hydrolaboratory), a full mock-up of the spacecraft, high-altitude barochambers, and a simulator of space suit system failure modes. The hydrolaboratory is typically used in constructing and testing various spacecraft components, apparatus, and crew equipment, and also for developing work techniques for the crew and determining what work operations can be performed during EVA within the limit loading values. Moreover, it provides wonderful opportunities for conducting preflight cosmonaut training. Here, cosmonauts become familiar with the schedule of planned operations and master skills of locomotion under conditions maximally simulating those of weightlessness. Before each submersion in the water, crew members undergo a brief medical examination and while they are submerged, they are subject to constant physiological monitoring.

Physiological criteria for EVA on long-term spaceflights include evaluating the possible effects of microgravity on crew performance efficacy. This means that every change in a physiological system, including changes in proprioception, strength and muscle mass, cardiopulmonary deconditioning, bone demineralization, and effects on vestibular functioning and gas exchange, may have some relation to the crew's readiness to perform an EVA effectively. For this reason, as part of EVA preparation, 2 weeks before the EVA, crew members undergo a comprehensive medical examination, including provocative tests on the bicycle ergometer.

During an EVA, the ground services, in addition to recording technical space suit parameters, monitor physiological parameters (EKG, pneumogram, body temperature, and caloric expenditure).

During and after an EVA, cosmonauts may develop various problems that require the attention of experts in space medicine. These may include

- general and local (midsized and small muscles of the arms) fatigue;
- pressure sores and blisters on the hands;
- emotional stress;
- shifts in thermal status (chilling, overheating);
- symptoms of high-altitude decompression sickness.

To prevent altitude decompression sickness brought on by the shift from the low working pressure inside the space suit (330 mmHg) to the normal atmospheric pressure within the spacecraft, immediately before the start of the EVA, crew members prebreathe pure oxygen at a pressure within the space suit of 533 mmHg for denitrogenation (20).

Postflight Medical Rehabilitation. Despite cosmonauts' use of measures provided by the prophylactic system, exposure to adverse spaceflight causes them to exhibit certain changes in the cardiovascular system's tolerance of the orthostatic position and in the characteristics of bone and muscle tissue, metabolic shifts, and vestibular and sensory impairment (see *Biological Responses and Adaptation to Spaceflight: Living in Space—An International Enterprise*), which require them to undergo medical rehabilitation postflight.

A system of rehabilitative and therapeutic measures has been adopted using data generated by clinical medicine, previous orbital flights, and ground-based simulation studies to help cosmonauts adapt to conditions on Earth, effectively restore their altered physiological functions and performance capacity, and ensure their professional longevity.

The specific rehabilitation program is based on

- the flight program (its duration and crew workload);
- characteristics of the crew-members' adaptation to spaceflight conditions;
- severity of fatigue (debilitation) during the flight;
- postflight changes in a cosmonaut's feeling of well-being and health status;
- individual characteristics of crew members and their preferences.

Structurally, the rehabilitation period consists of the following phases:

- meeting the crew at the landing site and evacuating it to a specialized rehabilitative and therapeutic base;
- readaptation at the rehabilitative-therapeutic base;
- recovery at a sanatorium or health spa.

The goal of the first phase is to provide a safe, nonstressful transition to conditions of normal gravity. This is achieved by limiting orthostatic, physical, and vestibular stress and through use of postflight prophylactic suits.

The most critical stage of readaptation, which takes place at a specialized rehabilitation and therapeutic base, lasts an average of 2–3 weeks. The major goal of this period is restoration of previous functional physiological status. During this phase, motor activity is gradually increased, and different techniques of rehabilitation (calisthenics, massage, and workouts in the pool and on exercise machines) are sequentially introduced. Loading is increased by increasing the rates of walking, running, or swimming, decreasing the duration and number of rest periods, and increasing the number of exercises and their difficulty.

Creating a favorable psychological climate and positive emotional factors are considered highly important. In addition, selecting and sequencing rehabilitation methods are based on the results of comprehensive batteries of postflight clinical and physiological tests.

The sanatorium/health spa phase of rehabilitation lasts for the next 20–30 days. Factors considered in site selection include climatic conditions at the time of year, level of equipment at the sanatorium, and the preferences of the crew members. During this stage, extensive use is made of climatic factors, physical therapy, and mud baths, the methods of therapeutic exercise and physical training, and long-distance running along a natural course. The rehabilitation and therapeutic methods employed are directed at completely restoring health status and functional physiological reserves. An individualized approach is used for prescribing procedures, accounting for health status and the temporal course of recovery and also of the capacities and desires of the crew members. Postflight psychological rehabilitation (the final phase of psychological support) is considered a very important part of this process. Such rehabilitation is directed at restoring social contacts, which have been partially lost (including family ties), as well as psychophysiological reserves that have been depleted under the extreme conditions of spaceflight (21).

Future Prospects for Developing the Biomedical Flight Support System

From a medical point of view, the use of the International Space Station (ISS) will be characterized by

- an increase in crew size and in the heterogeneity (including nationality) of the payload specialists involved;
- heightened work intensity, use of multiple shifts, an increase in the number of EVAs, and increased complexity of the EVA programs;
- permanent capacity to evacuate sick and injured from the ISS using specialized rescue spacecraft.

These characteristics will determine the set of specific medical, engineering psychological, and ergonomic tasks that must be performed, including

- developing criteria for a differential approach to selection and flight qualification for individuals who vary in initial health status, age, and gender;
- establishing criteria for permissible and optimal duration, the number of repeated flights, and the interval between them for various groups of cosmonauts;
- developing a system for medical support of rescue work;
- developing a system of measures for medical support of space crews whose members do not fully meet standards (for example, space tourists); this requires an individualized approach to regulating schedules of work, rest, meals, physical training and to conducting medical monitoring and therapeutic and preventive treatments;
- developing measures to support the safety of group EVAs for conducting planned work and rescue operations;

- developing a scientific basis for specifications of the ergonomic characteristics of the ISS, the structure and particular methods of medical support, and the functions of the crew physician.

Current attainments in this area are a good foundation for further progress in solving biomedical problems presented by future piloted space projects, including the Mars mission.

The isolation of the crew on a flight to Mars will require significantly more reliable safeguards both for the spacecraft systems and for the medical flight support system. The impossibility of emergency return of the crew to Earth or replacement of a sick crew member make it absolutely essential that a highly qualified physician-cosmonaut take part in the mission. An automated system for collecting, transmitting, and analyzing biomedical information and data banks and databases is already being created for the medical support of such autonomous spaceflights.

Another of the most important conditions to support a Mars mission is successful solution of a set of psychological problems, including psychological readiness to accept risk and to perform tasks at the limit of psychological and physical capacities and to resolve nonstandard contingency situations while living in isolation from Earth. An obligatory condition for inclusion in the crew must be a candidate's previous experience with long-term flights, because an interplanetary flight will mobilize all of the individual resources of the crew.

We will have to develop a more biologically complete and ecologically based artificial living environment fully appropriate to long-term human needs. We must create an analog of Earth's biosphere on board the Mars spacecraft, whose active components will be human beings, animals, plants, and microbes. Once this is accomplished, existing LSSs will be replaced by a regenerative LSS with a high coefficient of cycle closure. Laboratories on Earth have already produced encouraging results.

Another important problem is protection from galactic and solar cosmic radiation, which increase significantly outside the bounds of the radiation belts. On long-term interplanetary voyages, we will have to contend with the risk of mutagenic processes and also with threats to the lives and health of our cosmonauts. Approaches to ensuring radiation safety under these conditions may include selecting certain periods of solar activity for flights, creating a radiation shelter on board the spacecraft, and possibly, using pharmacological protective agents.

The risk of a Mars expedition is significantly higher than that associated with human presence in near-Earth orbit and thus interplanetary missions must be preceded by intensive in-depth research in the area of space physiology, psychology, radiobiology, life support systems, and the development of reliable means for protecting and maintaining the health of cosmonauts. For this purpose, along with the obligatory use of existing ground-based experimental bases and devices, we will need to make maximally effective use of the potential of existing and planned space stations and unpiloted biosatellites, which will allow us to create a strong foundation of global cosmonautics to conquer the planets of the solar system (22).

Glossary

ASE. Adverse Somatic Effect
BLSS. Biological Life Support System
EKG. Electrocardiogram
EVA. Extravehicular Activity
ISS. International Space Station
LSS. Life Support System
OS. Orbital Station
Sv. Sievert

BIBLIOGRAPHY

1. Bugrov, S.A., Yu.I. Voronkov, L.I. Voronin, et al. Selection and biomedical training of cosmonauts. *Adv. Space Res.* 12 (1): 347–350 (1992).
2. Voronkov, Yu.I., E.I. Mantsev, and M.P. Kuzmin. Some characteristics of medical selection of cosmonauts today. In F.P. Kosmolinskiy (ed.). *Collection on Space Biomedicine*. Tsiolkovsky Academy of Cosmonautics, Moscow, 1996, pp. 21–27.
3. Grigoriev, A.I., and A.D. Egorov. Theory and practice of medical monitoring on long-term space flights. *Aviakosmicheskaya i ekologicheskaya meditsina* 31 (1): 14–25 (1997).
4. Grigoriev, A.I., and A.D. Egorov. Medical monitoring in long-term space missions: Theory and experience. *The World Space Congress*, Washington, DC, August 28–September 5, 1992.
5. Bogomolov, V.V., I.B. Goncharov, and L.L. Stazhadze. Means and methods of medical care. In O.G. Gazenko (ed.), *Space Biology and Medicine: Physiological Handbook*. Nauka, Moscow, 1987, pp. 255–270.
6. Huntoon, C.S.L., A.I. Grigoriev, and Yu.V. Natochin. Fluid and electrolyte regulation in spaceflight. *Science and Technology Series*, 94. American Astronautical Society, Springfield, VA, 1998.
7. Kozlovskaya, I.B., A.I. Grigoriev, and V.I. Stepantsov. Countermeasures of the negative effects of weightlessness on physiological systems in long-term space flights. *Acta Astronautica* 36 (8/12): 661–668 (1995).
8. Guzenberg, A.S. Air regeneration in spacecraft cabins. In A.E. Nicogossian, S.R. Mohler, O.G. Gazenko, and A.I. Grigoriev (eds), *Space Biology and Medicine*, Joint U.S./Russian Publication in Five Volumes. AIAA, Washington DC, 1993, Vol. II, chap. 9, pp. 175–208.
9. Meleshko, G.I., and Ye.Ya. Shepelev. *Biological Life-Support Systems (Closed Ecological Systems)*. Sintez, Moscow, 1994.
10. Popov, I.G., and V.P. Bychkov. Crewmember nutrition. In A.E. Nicogossian, S.R. Mohler, O.G. Gazenko, and A.I. Grigoriev (eds), *Space Biology and Medicine*, Joint U.S./Russian Publication in Five Volumes. AIAA, Washington DC, 1993, Vol. II, chap. 11, pp. 223–238.
11. Sinyak, V.V. Gaidadmov, V.M. Skuratov, R.L. Sauer, and R.W. Murray. Spaceflight water supply. In A.E. Nicogossian, S.R. Mohler, O.G. Gazenko, and A.I. Grigoriev (eds), *Space Biology and Medicine*, Joint U.S./Russian Publication in Five Volumes, AIAA, Washington DC, 1993, Vol. II, chap. 12, pp. 239–264.
12. Nefedov, Yu.G. (ed.). *Sanitary Hygienic and Physiological Aspects of Inhabited Spacecraft*, Problems in Space Biology Volume 42. Nauka, Moscow, 1980.

13. Azhayeve, A.N., A.A. Berlin, G.A. Shumilina, J.D. Villarreal, and P.F. Grounds. Clothing and personal hygiene of space crewmembers. In A.E. Nicogossian, S.R. Mohler, O.G. Gazenko, and A.I. Grigoriev (eds), *Space Biology and Medicine*, Joint U.S./Russian Publication in Five Volumes, AIAA, Washington DC, 1993, Vol. II, chap. 6, pp. 125–138.
14. Popov, I.G., and V.P. Bychkov. Crewmember nutrition. In A.E. Nicogossian, S.R. Mohler, O.G. Gazenko, and A.I. Grigoriev (eds), *Space Biology and Medicine*, Joint U.S./Russian Publication in Five Volumes. AIAA, Washington DC, 1993, Vol. II, chap. 11, pp. 223–238.
15. Robins, J.E., V.M. Petrov, W. Schimmerling, and I.B. Ushakov. Ionizing radiation. In A.E. Nicogossian, S.R. Mohler, O.G. Gazenko, and A.I. Grigoriev (eds), *Space Biology and Medicine*, Joint U.S./Russian Publication in Five Volumes, AIAA, Washington DC, 1993, Vol. III, Book 2, chap. 17, pp. 155–205.
16. GOST (State Standard) 25645.215-85, Radiation safety of spacecraft crews during space flight. Safety standards for flights up to three years in duration. Moscow, 1986.
17. Myasnikova, V.I., and V.P. Salnitskiy (eds). *The Problem of Psychological Asthenization on Long Duration Space Flight*. SLOVO, Moscow, 2000.
18. Novikov, A.N., and K.V. Eskov. Group dynamics and crew interaction during isolation. *Space Biol. Med.* 5: 233–244 (1995).
19. Katountsev, V.P., Yu.Yu. Osipov, N.K. Gnoevaya, G.G. Tarasenkova, and A.S. Barer. The main results of EVA medical support on Mir space station, *Int. Astronaut. Congr.*, Toulouse, Oct. 1–4 (2001).
20. Barer, A.S., and S.N. Filipenkov. Performance of suited cosmonauts. In O.G. Gazenko (ed.), *Space Biology and Medicine: Physiological Handbook*. Nauka, Moscow, 1987, chap. 5, pp. 146–176.
21. Vasilyeva, T.D., and B.M. Fedorov. Recovery of cosmonaut status postflight. In O.G. Gazenko (ed.), *Space Biology and Medicine: Physiological Handbook*. Nauka, Moscow, 1987, pp. 270–285.
22. Grigoriev, A.I., A.D. Egorov, and A.N. Potapov. Some medical problems relating to a manned Mars mission. *Aviakosmicheskaya i Ekologicheskaya Meditsina* 34 (3): 6–12 (2000).

ANATOLY I. GRIGORIEV
DMITRY K. MALASHENKOV
Institute of Biomedical Problems
Russian Academy of Sciences
Moscow, Russia

C

CARDIOVASCULAR SYSTEM IN SPACE

The microgravity environment of space is well tolerated by the cardiovascular system over during periods ranging from weeks to several months. In many ways, the heart, peripheral vasculature, and central cardiovascular control system are exposed to fewer challenges in microgravity than on Earth. Consider the simple act of standing upright in the normal gravity field of Earth. Assumption of an upright posture on Earth causes a redistribution of blood volume to the lower parts of the body, resulting in an increase in blood pressure in dependent blood vessels and a decrease in arterial blood pressure above the level of the heart. The cardiovascular system must quickly compensate for these changes by altering the heart rate, the force of contraction of the heart, and the resistance of the blood vessels to maintain enough blood flow to the brain to prevent loss of consciousness. Under weightless conditions, however, there are no postural changes in blood volume and pressure, which greatly reduces the demands placed on the cardiovascular system to maintain homeostasis. Furthermore, moving about in microgravity requires far less energy than required in gravity, which places less demand on the cardiovascular system.

During the long term, however, many physiological systems become dysfunctional when they are not required to perform at a normal level. For example, the muscles and bones of a leg immobilized in a cast for a month or two become atrophied and weak. Similarly, the relatively unchallenging environment of microgravity ultimately results in dysfunctional changes in the heart, the vasculature, and the central cardiovascular control system that may have harmful consequences during spaceflight and upon reexposure of the human organism to a gravitational field.

For example, many astronauts cannot maintain normal blood pressure and feel dizzy or faint when standing upright immediately following spaceflight, a

condition called *orthostatic intolerance*. This may impair their ability to get out of a spacecraft quickly should an emergency arise, or to perform meaningful work upon arrival in a gravitational field, for example, after a trip to Mars. There are also reports of rhythmic disturbances of the heart (*cardiac arrhythmias*) during spaceflight and of loss of muscle mass of the heart (*cardiac atrophy*). The latter two conditions, though less well documented than orthostatic intolerance, represent potentially life-threatening alterations in cardiovascular function. Before examining these problems in more detail, however, we will present a brief review of normal cardiovascular physiology.

Cardiovascular Physiology

The heart is composed of four chambers, the left and right atria and the right and left ventricles. The atria serve as booster pumps to aid in filling the ventricles during their filling cycle—ventricular diastole. The ventricles are the main pumping chambers of the heart. Blood is ejected from the ventricles during their contraction cycle—ventricular systole. The right ventricle pumps blood to the lungs through the pulmonary circulation. The left ventricle pumps blood through the systemic circulation. The systemic circulation is a branching network of vessels. The arteries bring blood to the various tissue beds. Oxygen and nutrients are delivered to the tissues, and carbon dioxide and waste products are removed through the smallest vessels, the capillaries. Capillary blood flows into the venous system starting with the smallest vessels in the venous system, the venules. The smaller veins merge into progressively larger veins ending up in the vena cava which returns blood back to the heart.

From a functional point of view, the large arteries serve as compliant capacitance vessels, ensuring that blood pressure does not fall to zero during ventricular diastole during which no blood is being ejected from the ventricles. The bulk of the resistance to flow in the systemic circulation resides in the microcirculation, consisting of arterioles, capillaries, and venules. The control of resistance to flow resides in the arterioles which, unlike capillaries and venules, have a muscular wall whose tone can be controlled by local factors (autacoids), by circulating hormones, and by the sympathetic nervous system. The large veins are extremely compliant and serve as a large blood reservoir. These veins have muscular walls whose tone can be increased by sympathetic nervous system stimulation. Constricting the large veins functionally is equivalent to shifting blood from the reservoir into the remainder of the circulation. This is an important function because filling of the right ventricle is determined to a large extent by the pressure in the vena cava—the central venous pressure—which may also be called *preload*. A drop in central venous pressure can lead to a precipitous drop in the filling of the right ventricle which in turn leads to a drop in cardiac output—the total rate of blood flow out of the heart into the systemic circulation. The large veins also provide small resistance to blood flow.

The cardiovascular system is exquisitely controlled by multiple feedback and control loops. Intrinsic control of the cardiovascular system is achieved by local factors (autacoids) that, for example, control the muscular tone in arterioles to match local blood flow to local tissue demand. Extrinsic control of the

cardiovascular system is achieved by the autonomic nervous system and circulating hormones. The autonomic nervous system is composed of two main branches—the parasympathetic nervous system and the sympathetic nervous system. The main function of the parasympathetic nervous system is to slow the heart rate. The sympathetic nervous system has two classes of receptors, alpha and beta. Stimulation of alpha receptors increases venous and arterial tone. Stimulation of beta receptors increases the heart rate and the contracting force of the heart (inotropy) and decreases arteriolar tone.

Cardiovascular Alterations Associated with Spaceflight

The constellation of cardiovascular deconditioning effects associated with spaceflight include decreased orthostatic tolerance (1–8) and exercise capacity (9,10) upon return to a gravitational field, decreased cardiac muscle mass (11), and the occurrence of a variety of arrhythmias in some individuals (12–14). Maintaining exercise capacity and orthostatic tolerance at preflight levels requires the integrity of both cardiac pump function and the multiple neurohumoral control mechanisms that mediate the hemodynamic response to exercise and orthostatic challenge. Orthostatic intolerance is a high priority problem because it may interfere with the crew's ability to function during reentry and postflight; therefore, we will start our discussion with this problem. In later sections, we will discuss cardiac arrhythmias and cardiac atrophy.

Alterations in Cardiovascular Parameters That May Contribute to the Development of Orthostatic Intolerance. Exposure to microgravity undoubtedly removes the blood pressure gradients from head to feet that are associated with upright posture on Earth (15). Thus, there is an equalization in blood pressures throughout the body. Mean arterial pressure at the feet is reduced from about 200 to about 100 mmHg and is increased within the head from about 70 to about 100 mmHg. Dependent blood vessels are exposed to lower than 1G normal blood pressure, whereas the vessels between the heart and head are exposed to higher than 1G normal blood pressure.

During spaceflight, body fluid shifts from the lower extremities to the thorax, and overall fluid volume is reduced. It is believed that the mechanism of orthostatic hypotension following spaceflight involves pooling of blood in the legs resulting in reduced preload to the heart, a decrease in cardiac output, and low blood pressure (*hypotension*). About 20% of astronauts after short (weeks) missions and 83% of astronauts after long missions (months) cannot support standing arterial blood pressure for 10 minutes (8). Physiological mechanisms that contribute to orthostatic hypotension include alterations in peripheral vascular resistance, venous compliance, intrinsic vascular reactivity, reduced intravascular volume, altered heart rate arterial baroreflex, and altered cardiac systolic and diastolic function.

Altered Vascular Resistance

Changes in Total Peripheral Resistance (TPR). Fourteen crew members from the Space Life Sciences (SLS)-1 and -2 Space Shuttle flights were studied within 4 hours of landing after flights of 9 or 14 days (3). Hemodynamic measurements were compared between finishers and nonfinishers of a 10-minute

stand test. Only nine of the 14 subjects (64%) finished the stand test. There were equally significant postflight increases in upright heart rates and decreases in stroke volumes in both finishers and nonfinishers. The amount of venous pooling was similar in both groups. The critical difference between finishers and nonfinishers in this series was inadequate vasoconstrictor response in nonfinishers (29.4 ± 2.3 units in finishers vs. 19.9 ± 1.4 units in nonfinishers, $p < 0.05$). Although the vasoconstrictor response increased compared to preflight levels in both subgroups, only the finishers had vasomotor responses enhanced enough to maintain adequate arterial blood pressure. These investigators concluded that microgravity-induced hypovolemia is a likely prerequisite for developing postflight orthostatic intolerance, but the outcome in a given individual may depend to a great extent on the magnitude of the systemic vasoconstrictor response.

Although the mechanisms that underlie inadequate postflight response remain to be identified, two principal alternatives are (1) adaptation to microgravity has caused a degradation of neurohumoral cardiovascular control mechanisms that are essential at 1G, or (2) the dynamic range of the mechanisms that produce appropriate orthostatic vasoconstriction is an inborn characteristic of the individual. A limited range that is adequate for an ordinary 1G condition becomes inadequate in the hypovolemic state early after return from space. A degradation of the neurohumoral vasoconstrictor mechanisms may occur at one or more levels, that is, afferent input, central integration, efferent output, and/or end organ responsiveness. Currently available information provides no conclusive answers (16).

In a study of 24 astronauts before and after missions of 4 to 5 days using two-dimensional echocardiography, the standing TPR index (TPRI) was significantly greater ($p < 0.03$) in the standing compared with the supine position on all test days except landing day. Similarly, the TPRI orthostatic response decreased on landing day ($p < 0.03$). Thus, there was an apparent reduction in the ability to augment peripheral vascular tone when assuming the standing position (4).

In a study of 40 astronauts before and after spaceflights that lasted up to 16 days, it was found that those who could not complete a 10-minute stand test on landing day had significantly lower (23 ± 3 units vs. 34 ± 3 units; $p = 0.02$) standing TPR (7).

Changes in Levels of Catecholamines. Forty astronauts were studied before and after spaceflights of up to 16 days. Of the original 40, seven were excluded because they had consumed promethazine, dextroamphetamine, or caffeine shortly before landing, and four were excluded because blood samples were ruined. On landing day, eight of the remaining 29 astronauts (28%), could not complete a 10-minute stand test due to presyncopal symptoms (dizziness or faintness). It was found that those who did not complete the stand test had significantly reduced peripheral vascular resistance and blood pressure when standing. These same subjects, it was noted, had significantly lower peripheral vascular resistance and supine and standing diastolic and systolic blood pressure before spaceflight. In addition, the nonfinishers had significantly smaller increases in plasma norepinephrine levels when standing than those who finished the stand test (105 ± 41 vs. 340 ± 62 pg/ml; $p = 0.05$). These results were taken as evidence for hypoadrenergic responsiveness, possibly centrally mediated, as a contributing factor in postflight orthostatic intolerance (7).

A 2-week head-down bed-rest study (a ground-based model of weightlessness) of eight healthy volunteers demonstrated a decrease in norepinephrine excretion of 35% on day 14 of bed rest from that on the control day. Though excretion rates of norepinephrine decreased, plasma levels were only variably and not significantly decreased. This was likely to be related to concurrent hypovolemia and is still consistent with a subnormal norepinephrine spillover in this setting. Excretion rates of epinephrine, dopamine, and dihydroxyphenylacetic acid were unchanged, suggesting that head-down bed rest produces sustained inhibition of sympathoneural release, turnover, and synthesis of norepinephrine without affecting adrenomedullary secretion or renal dopamine production. This sympathoinhibition in the face of decreased blood volume may help to explain orthostatic intolerance in returning astronauts (17).

The discordance between the response of plasma dopamine and plasma norepinephrine documented above may further exacerbate the deleterious effects of the sympathoinhibition. Renal tubular cells can synthesize dopamine from dopa (18), and in normal circumstances, the diuretic, natriuretic, and renal vasodilatory effects of locally produced dopamine are balanced by the effects of renal sympathetic nerve activity and the resultant release of norepinephrine, which promotes renal absorption of sodium and water and reduces dopamine-induced vasodilation (19). It is interesting to note that an extreme example of discordance between plasma dopamine and plasma norepinephrine is found in patients who are deficient in the enzyme dopamine beta-hydroxylase (20). In these patients, excessive production of dopamine, coupled with an inability to convert dopamine to norepinephrine, leads to a volume-depleted state with extraordinarily severe orthostatic hypotension, which is enhanced by the absence of the vasoconstricting properties of norepinephrine.

Other studies have shown a dissociation between an increase in circulating catecholamines and peripheral vasoconstrictor responses, from which it was concluded that there is a blunted vasoconstrictor response to sympathetic stimulation (5,21) following exposure to microgravity. This may be due to downregulation of adrenergic receptors in response to increased levels of plasma norepinephrine in microgravity. During the D2-Spacelab mission, plasma norepinephrine in four astronauts was approximately twice the value of that in the supine position on the ground, suggesting that the level of sympathetic nervous activity during microgravity is more similar to the upright ground-based position than to the supine (22).

Changes in Local Mediators. Endothelial functional changes resulting from exposure to microgravity have yet to be extensively studied. As outlined before, some studies document a blunted vasoconstrictor response to sympathoadrenal activation after spaceflight (5,21), whereas others have shown sympathoinhibition (7,17). In either case, it is likely that local endothelial vasodilatory function mediated by endothelium-derived relaxing factor, or nitric oxide, is altered following exposure to microgravity.

Altered Venous Compliance

Central Venous Pressure (CVP). There is some evidence that central venous pressure is decreased during spaceflight. CVP was measured in one subject aboard Space Life Sciences-1 and in two subjects aboard Space Life Sciences-2. Mean CVP in the seated position prelaunch was 8.4 cm water, and with legs

elevated, prelaunch in the Shuttle was 15.0 cm water and fell to 2.5 cm water after 10 minutes in microgravity. In these same subjects, however, the left ventricular end-diastolic dimension, as measured by echocardiography, increased within 48 hours in microgravity (23). Given this increase in cardiac filling, it seems likely that a decrease in intrathoracic pressure of greater magnitude than the decrease in CVP may occur in weightlessness, leading to an effective increase in right atrial transmural pressure. Simultaneous measurements of CVP and intraesophageal pressure, recently made in weightless parabolic flight, confirm this hypothesis (24). The SLS-1 and -2 data confirm previous observations in space (25).

Increased Peripheral Pooling of Blood. Data from Skylab-3 and -4 suggest that leg blood flow and compliance increase during the early hours of spaceflight (26), and ground-based studies simulating microgravity have documented an increase in leg vessel compliance (27–29).

Reduced Intravascular Volume

Decreased Red Blood Cell Mass (RBCM). Six astronauts on SLS-1 and -2 were studied. Plasma volume (PV) decreased by 17% within the first day of spaceflight. RBCM decreased as a result of the destruction of red blood cells (RBCs) either newly released or about to be released from the bone marrow, whereas older RBCs survived normally. Upon return to Earth, PV increased, causing a decreased RBC count and increased erythropoietin levels. The proposed mechanism for these changes is that entry into microgravity causes acute plethora secondary to a decreased vascular space, leading to increased hemoglobin, that leads to decreased erythropoietin levels. RBCM decreases to an appropriate level for the microgravity-induced decreased PV via destruction of recently formed RBCs. Acute hypovolemia upon return to Earth stimulates an increase in plasma volume, leading to anemia that stimulates an increase in serum erythropoietin and corrects the anemia (30).

Decreased Plasma Volume (PV). Adaptation to actual and simulated microgravity is associated with decreased total blood volume (2,31). It was initially postulated that diuresis accounts for fluid volume losses during spaceflight; however, diuresis during spaceflight has rarely been documented (32). Body fluid balance was studied on three recent spaceflights (N of subjects = 7) with special emphasis on oral intake and renal excretion of fluid and sodium (33). In no case was increased diuresis and natriuresis observed; rather, both oral fluid and sodium intake, as well as renal fluid and sodium output, appear reduced compared with the preflight condition. These results are consistent with findings during the Skylab program, in which fluid intake and renal fluid loss also appeared reduced or unchanged, at least during the first flight days (34). They are also consistent with the results of SLS-1 and -2 subjects in whom plasma volume, extracellular fluid volume, urine excretion, and fluid intake all decreased, and the glomerular filtration rate (GFR) was elevated (32). It was felt that the reason for the discrepancy between reduced or unaltered renal excretion and reduced body water was most likely due to insufficient caloric intake and subsequent decreased water binding capacity (1 g glycogen binds 3–4 g water; 1 g protein binds 8 g water). In addition, the decreased plasma volume coupled with upper body edema points toward an increased extravasation of fluid as a result of the headward fluid shift (33). Leach et al. postulate that increased permeability of

capillary membranes may be the most important mechanism causing spaceflight-induced PV reduction, which is probably maintained by increased GFR and other mechanisms (32).

Hormonal Changes. Following an isotonic saline infusion in microgravity, renal sodium and fluid output were lower than expected from results of simulation experiments; venous plasma norepinephrine and renin were higher. Because plasma arginine vasopressin (AVP) was low, high levels of this peptide were not responsible for decreased renal fluid output during flight (22,35).

SLS-1 and -2 subjects had increased AVP (also referred to as antidiuretic hormone, or ADH) on flight day 1 and on landing day; AVP levels normalized on other days. The elevations on launch and landing days, it was felt, were stress related. Plasma and urinary cortisol levels were elevated, although not statistically significant, throughout the flights, and again, it was felt that stress plays a role in this elevation. Plasma renin activity (PRA) and aldosterone decreased in the first few hours after launch, but PRA was elevated 1 week later. During flight atrial natriuretic peptide concentrations were consistently lower than preflight mean values, (32).

Altered Cardiac Function

Changes in Heart Rate (HR) and the HR Baroreflex. Using a neck collar to produce computer-controlled beat to beat changes in carotid artery transmural pressures in subjects before and after 8 days in orbit in SLS-1 and D2 flights demonstrated a significantly attenuated change in the cardiac cycle length interval for a given change in carotid transmural pressure (36). Similar results were found after head-down bed rest (37). The conclusion of these studies was that spaceflight reduces baseline levels of vagal-cardiac outflow and vagally mediated responses to changes of arterial baroreceptor input. These results were recently corroborated by studies of the spectral power of heart rate and blood pressure in Mir cosmonauts during and after 9 months of spaceflight (38). It has been argued, however (15), that baroreflex assessments based on R-R interval changes alone may be inadequate because they ignore the contribution of stroke volume to cardiac output. For example, HR increases more during in-flight lower body negative pressure (LBNP) than during preflight LBNP, and standing HR is elevated postflight to compensate for the stroke volume deficit imposed by microgravity-induced hypovolemia. These observations alone imply adequate functioning of the cardiac baroreflex arm.

A study of 24 astronauts following missions from 4 to 5 days found that, on landing day, supine HR increased by 23% ($p < 0.0005$) and standing HR increased by 35% ($p < 0.0001$), compared with preflight values. Preflight HR began to level off 2 or 3 minutes following the initial increase when standing, but postflight, it continued to increase for the duration of the 5-minute stand test. This was taken as evidence of postflight orthostatic dysfunction (4,6).

Changes in Ventricular Filling and Cardiac Contractility. After Apollo flights, standard anterior-posterior chest roentgenograms taken pre- and postflight showed that heart size decreased following spaceflight (39). These results have been confirmed in a study of 24 astronauts after missions of 4 to 5 days evaluated by two-dimensional echocardiography. Supine left ventricular end diastolic volume index (EDVI) diminished by 11% ($p < 0.04$) on landing day, compared with preflight. Supine left ventricular stroke volume index (SVI)

diminished by 17% ($p < 0.006$) on landing day, compared with preflight. Both recovered to preflight levels within 48 hours. Standing EDVI was less than that of supine EDVI, but SVI did not change significantly with position. The ejection fraction and the velocity of circumferential fiber shortening did not change significantly, suggesting no effect on myocardial contractility (4).

More recent early in-flight echocardiographic measurements in three subjects from SLS-1 and -2 showed an increase in cardiac filling, the mean increase in the left ventricular (LV) diastolic diameter was from 4.6 cm to 4.97 cm (23). Analyzing the data by a technique that produces a three-dimensional reconstruction of the LV showed a time course of adaptation where the initial increase in LV size was followed within 48 hours by a significant decrease in size relative to preflight supine dimensions (16). Contractile state—as defined by the LV ejection fraction, end systolic volume, and velocity of circumferential fiber shortening—did not change during the mission. Stroke volume measurements after 2 days in space approximated the 1 G supine data, and measurements after 5 days or later approached, but did not reach, preflight upright levels. This suggests that cardiovascular conditions in microgravity after adaptation may represent an intermediate hemodynamic state that accurately reflects the normal 24-hour human postural pattern, that is, one 8 hour part supine and two parts upright (16,23).

A recent bed-rest study suggests that changes in LV pressure–volume characteristics may develop during a 2-week period of bed rest in normal subjects and produce a stiffer ventricle that has reduced end-diastolic volume, compared to the normal physiological range of filling pressures (9).

Alterations in Cardiac Muscle Mass. Magnetic resonance imaging of four members of the German D-2 German Spacelab mission showed a significant loss of myocardial mass after a 10-day mission (11). Levine et al. (40) found that normal subjects subjected to 2 weeks of microgravity simulated by bed rest showed a significant reduction in ventricular mass also, as measured by magnetic resonance imaging. However, the mechanisms by which reduction in cardiac mass occurs, the functional sequelae of the changes in cardiac mass, and whether or not these changes are reversible after space flight all remain to be investigated.

Alterations in Cardiac Electrical Function. Another aspect of the cardiovascular deconditioning process involves potential alterations of cardiac conduction processes associated with spaceflight. Cardiovascular deconditioning has been investigated quite extensively, but there have been relatively fewer systematic investigations into the effects of spaceflight on cardiac electrical function. However, a variety of heart rhythm disturbances have been observed in astronauts during and after spaceflight. Occasional premature ventricular contractions were seen in Gemini and Apollo missions (13,41). Reports indicate that all crew members in the Skylab series had some form of rhythmic disturbance (14,41) and one individual experienced a five-beat run of ventricular tachycardia. The incidence of arrhythmias was higher during flight than during preflight testing and higher than would be expected in a random sampling of a healthy population. Cardiac arrhythmias have also been seen during Shuttle flights (41) and on Mir. Analysis of nine 24-hour ECG recordings (Holter monitoring) obtained during long-term spaceflight on Mir revealed one 14-beat run of ventricular tachycardia (12). Two Mir missions have undergone major changes in crew composition and/or responsibilities due to cardiac dysrhythmias (42). Fur-

thermore, a research primate recently died suddenly shortly after returning to Earth from extended space flight; cardiac dysrhythmic mechanisms were suspected as a possible cause (43).

Thus, there seems to be significant anecdotal evidence suggesting that spaceflight is associated with an increased susceptibility to potentially life-threatening ventricular arrhythmias. Furthermore, it is likely that ventricular arrhythmias during spaceflight will be of increasing concern in the future as older individuals are involved in spaceflight and as the durations of missions lengthen. At a joint National Aeronautics and Space Administration/National Space Biomedical Research Institute workshop in January 1998, cardiac arrhythmias were identified as the leading cardiovascular risk to a human Mars exploration mission (44). Older individuals have a greater statistical likelihood of having underlying structural heart disease, in particular, coronary artery disease and thus, will be at greater risk for heart rhythm disturbances. If spaceflight does increase susceptibility to ventricular arrhythmias, such arrhythmias could pose a significant threat to crew safety and mission success. However, the available data are too anecdotal to permit one to conclude whether spaceflight does increase susceptibility to ventricular arrhythmias. Therefore, it is important to conduct systematic investigations to determine whether exposure to microgravity alters cardiac electrical stability.

Recently a new technique—the measurement of microvolt level T wave alternans (TWA)—has been developed. In a series of clinical studies in varied patient populations, this technique compared favorably to other noninvasive risk stratifiers and invasive electrophysiological testing as a predictor of sudden cardiac death, ventricular tachycardia, and ventricular fibrillation (45–47). In a recent National Space Biomedical Research Institute project, healthy volunteers participated in a 16-day head-down, tilt, bed-rest study (a ground-based analog of weightlessness) and had TWA measured before and after the bed-rest period during bicycle exercise stress. In three subjects, bed rest induced sustained TWA, although they had an onset heart rate above the 110 beat per minute (bpm) cutoff below which TWA is clinically associated with increased arrhythmic risk. In these subjects, sustained TWA disappeared 2 to 3 days after bed rest. In one subject who had sustained TWA and an onset heart rate above 110 bpm before bed rest, bed rest abolished sustained TWA, which reappeared 3 days after bed rest. These findings provide the first evidence that simulated weightlessness has a measurable effect on myocardial repolarization processes, which suggests that spaceflight may alter susceptibility to life-threatening ventricular arrhythmias (48).

The potential lethal arrhythmic risk for astronauts is sustained ventricular tachycardia or ventricular fibrillation. Nonsustained ventricular tachycardia could cause syncope. Given the data suggesting that cardiac arrhythmias might pose a problem for long-term spaceflight and given that the consequence of ventricular arrhythmias may be astronaut death, this will be an important area of further study.

Countermeasures

Fully effective countermeasures against the problem of orthostatic intolerance have yet to be developed. Currently, the US space program regularly employs

anti-g suits for reentry to limit acutely the amount of blood pooling in the lower extremity and oral intake of saline before reentry in an attempt to restore intravascular volume closer to preflight levels. Oral saline loading appears to be effective in preserving standing arterial blood pressure after short flights of up to 5 days (49); however, it has not been as effective after longer flights. This finding is consistent with other evidence that changes in autonomic control of cardiovascular function, and perhaps myocardial contractile function, rather than simply inadequate intravascular volume, is responsible for orthostatic intolerance following spaceflight.

One technique that appeared promising in the past was the use of lower body negative pressure (LBNP) during flight to simulate exposure periodically to gravity gradients. The LBNP device is a cylinder into which the subject's lower body is placed, and it has a rubber cuff that goes around the waist. A partial vacuum created in the cylinder causes blood to pool in the legs, similar to standing in Earth's gravity. Although exposure to LBNP during flight was moderately protective against orthostatic intolerance upon return to Earth, it was unpopular with astronauts, who reported that exposure to LBNP during flight caused them to reexperience many of the unpleasant sensations they had experienced upon initial exposure to weightlessness (50), such as nasal stuffiness and facial fullness. This is consistent with the idea that the cardiovascular system makes appropriate adaptations to weightlessness and that trying to simulate gravitational conditions periodically during flight may interfere with these appropriate inflight adaptations.

A new pharmacological countermeasure, midodrine, was recently tested and found protective against orthostatic intolerance following microgravitational simulation for 16 days (51). There is evidence that both venous return and peripheral vascular resistance are reduced after spaceflight. Though not the only contributors, both of these factors most certainly increase the incidence of post-spaceflight orthostatic hypotension and presyncope. Several studies have demonstrated a reduction in cardiac stroke volume upon return from space (3,4,7), and others have shown reduced resistance responses to standing, particularly in those astronauts who have the most difficulty maintaining arterial blood pressure while standing (3,7). Midodrine is an agonist at α -adrenergic receptors located on the smooth muscle in both veins and arterioles, and thus reduces venous pooling and increases peripheral vascular resistance (52–54) by reducing the diameters of arterioles and veins. The success of this trial after bed rest (51) suggests that midodrine also may be an effective treatment for orthostatic hypotension following spaceflight. Trials of midodrine following spaceflight are now being conducted.

Conclusion

The cardiovascular system performs fairly well during spaceflight. The primary adverse effects of spaceflight on the cardiovascular system are orthostatic intolerance, cardiac arrhythmias, and cardiac atrophy. Spaceflight also impairs exercise intolerance, but current in-flight exercise programs seem to be an adequate countermeasure. Preexisting silent cardiovascular disease (such as

coronary artery disease) may also become manifest during long-duration spaceflight, particularly in older astronauts.

In terms of the primary adverse effects, postflight orthostatic intolerance is a current operational problem that occurs at high frequency, particularly after long-duration spaceflight. Orthostatic intolerance is generally not life-threatening. The task here is to develop and test effective countermeasures.

It is not known whether or not spaceflight increases the risk of life-threatening cardiac arrhythmias. However, if it does, the consequences could be the death of an astronaut. The task here is to establish whether or not spaceflight increases susceptibility of the heart to life-threatening arrhythmias and if so, to develop effective countermeasures.

Finally, it appears that short-duration spaceflight may modestly decrease cardiac mass. It is not known whether the degree of cardiac atrophy increases with the duration of spaceflight, what are the mechanisms involved, what are the functional sequelae, and whether this effect is reversible after spaceflight. The task here is to answer these questions and if indicated, develop effective countermeasures.

BIBLIOGRAPHY

1. Blomqvist, C.G. Orthostatic intolerance. In W.W. Parmley and K. Chatterjee (eds), *Cardiology*, Lippincott, Philadelphia, PA, 1990, pp. 1–37.
2. Blomqvist, C.G., and H.L. Stone. Cardiovascular adjustments to gravitational stress. In *Handbook of Physiology. The Cardiovascular System. Peripheral Circulation and Organ Blood Flow*. American Physiological Society Bethesda, MD, 1983, Sect. 2, vol. III, pt. 2, Chap. 28, pp. 1025–1063.
3. Buckey, J.C., L.D. Lane, B.D. Levine, D.E. Watenpaugh, S.J. Wright, W.E. Moore, F.A. Gaffney, and C.G. Blomqvist. Orthostatic intolerance after spaceflight. *J. Appl. Physiol.* 81(1): 7–18 (1996).
4. Mulvagh, S.L., J.B. Charles, J.M. Riddle, T.L. Rehbein, and M.W. Bungo. Echocardiographic evaluation of the cardiovascular effects of short duration space flight. *J. Clin. Pharmacol.* 31: 1024–1026 (1991).
5. Fritsch-Yelle, J.M., J.B. Charles, M.M. Jones, L.A. Beightol, and D.L. Eckberg. Spaceflight alters autonomic regulation of arterial pressure in humans. *J. Appl. Physiol.* 77: 1776–1783 (1994).
6. Charles, J.B., and C.M. Lathers. Cardiovascular adaptation to spaceflight. *J. Clin. Pharmacol.* 31: 1010–1023 (1991).
7. Fritsch-Yelle, J.M., P.A. Whitson, R.L. Bondar, and T.E. Brown. Subnormal nor-epinephrine release relates to presyncope in astronauts after spaceflight. *J. Appl. Physiol.* 81(5): 2134–2141 (1996).
8. Reyes, C., S. Perez, and J. Fritsch-Yelle. Orthostatic intolerance following short and long duration spaceflight. *FASEB J.* 13(5): A1048 (1999).
9. Levine, B.D., J.H. Zuckerman, and J.A. Pawelczyk. Cardiac atrophy after bed-rest deconditioning: A non-neural mechanism for orthostatic intolerance. *Circulation* 96(2): 517–525 (1997).
10. Shykoff, B.E., L.E. Farhi, A.J. Olszowka, D.R. Pendergast, M.A. Rokitka, C.G. Eisenhardt, and R.A. Morin. Cardiovascular response to submaximal exercise in sustained microgravity. *J. Appl. Physiol.* 81(1): 26–32 (1996).

11. Blomqvist, C.G., et al. In *Scientific Results of the German Spacelab Mission D-2, Proceedings of Symposium at Norderney*, P.R. Sahm et al. (eds). Wissenschaftliche Projektführung D-2 RWTH Aachen, care of DLR, Köln, pp. 688–690.
12. Fritsch-Yelle, J.M., U.A. Leuenberger, D.S. D'Aunno, et al. An episode of ventricular tachycardia during long-duration spaceflight. *Am. J. Cardiol.* 81(11): 1391–1392 (1998).
13. Hawkins, W.R., and J.F. Zieglschmid. Clinical aspects of crew health. In *Biomedical Results of Apollo* (NASA SP-368), R.S. Johnston, L.F. Dietlein, and C.A. Berry (eds). U.S. Government Printing Office, Washington, DC, 1975, pp. 43–81.
14. Smith, R.F., K. Stanton, D. Stoop, D. Brown, W. Januez, and P. King. Vectorcardiographic changes during extended space flight (M093): Observations at rest and during exercise. In *Biomedical Results of Skylab* (NASA SP-377), R.S. Johnston, and L.F. Dietlein (eds). NASA, Washington, DC, 1977, pp. 339–350.
15. Hargens, A.R. and D.E. Watenpugh. Cardiovascular adaptation to spaceflight. *Med. Sci. Sports Exercise* 28(8): 977–982 (1996).
16. Blomqvist, C.G. Regulation of the systemic circulation at microgravity and during readaptation to 1G. *Med. Sci. Sports Exercise* 28 (10, Suppl.) S9–S13 (1996).
17. Goldstein, D.S., J. Vernikos, C. Holmes, and V.A. Convertino. Catecholaminergic effects of prolonged head-down bed rest. *J. Appl. Physiol.* 78(3): 1023–1029 (1995).
18. Hayashi, M., Y. Yamaji, W. Kitajima, and T. Saruta. Aromatic L-amino acid decarboxylase activity along the rat nephron. *Am. J. Physiol.* 258: F28–33 (1990).
19. DiBona, G.F., and C.S. Wilcox. The kidney and the sympathetic nervous system. In *Autonomic Failure* R. Bannister, and C.J. Mathias (eds). Oxford University Press, Oxford, 1992, pp. 178–196.
20. Robertson, D., S.E. Perry, A.S. Hollister, R.M. Robertson, and I. Biaggioni. Dopamine beta-hydroxylase deficiency: A genetic disorder of cardiovascular regulation. *Hypertension* 18: 1–8 (1991).
21. Whitson, P.A., J.B. Charles, W.J. Williams, and N.M. Cintron. Changes in sympatho-adrenal response to standing in humans after spaceflight. *J. Appl. Physiol.* 79: 428–433 (1995).
22. Norsk, P., C. Drummer, and L. Rocker. Renal and endocrine responses in humans to isotonic saline infusion during microgravity. *J. Appl. Physiol.* 264: 2253–2259 (1995).
23. Buckey, J.C., F.A. Gaffney, L.D. Lane, B.D. Levine, D.E. Watenpugh, S.J. Wright, C.W. Yancy, D.M. Meyer, and C.G. Blomqvist. Central venous pressure in space. *J. Appl. Physiol.* 81(1): 19–25 (1996).
24. Pantalos, G., et al. Venous and esophageal pressure in humans during parabolic flight. *FASEB J.* 13(4): A108 (1999).
25. Kirsch, K.A., L. Rocker, O.H. Gauer, R. Krause, C. Leach, H.J. Wicke, and R. Landry. Venous pressure in weightlessness. *Science* 255: 218–219 (1984).
26. Thornton, W.E., G.W. Hoffer, and J.A. Rummel. Hemodynamic studies of the legs under weightlessness. In: *Biomedical Results from Skylab*, edited by R.S. Johnston and L.F. Dietlein (eds). National Aeronautics and Space Administration, Washington, DC, 1977, pp. 324–329.
27. Beck, L., F. Baisch, F.A. Gaffney, J.C. Buckey, P. Arbeille, F. Patat, A.D.J. Ten Harkel, A. Hillebrecht, H. Schulz, J.M. Karemaker, M. Meyer, and C.G. Blomqvist. Cardiovascular response to lower body negative pressure before, during, and after ten days head-down bedrest. *Acta. Physiol. Scand.* 144 (Suppl. 604): 43–52 (1992).
28. Buckey, J.C., L.D. Lane, G. Plath, F.A. Gaffney, F. Baisch, and C.G. Blomqvist. Effects of head-down tilt for 10 days on the compliance of the leg. *Acta. Physiol. Scand.* 144 (Suppl. 604): 53–60 (1992).
29. Convertino, V.A., D.F. Doerr, and S.L. Stein. Changes in the size and compliance of the calf after 30 days of simulated microgravity. *J. Appl. Physiol.* 66: 1509–1512 (1989).

30. Alfrey, C.P., M.M. Udden, C. Leach-Huntoon, T. Driscoll, and M.H. Pickett. Control of red blood cell mass in spaceflight. *J. Appl. Physiol.* 81(1): 98–104 (1996).
31. Johnson, P.C., T.B. Driscoll, and A.D. LeBlanc. Blood volume changes. In R.S. Johnston and L.F. Dietlein (eds). *Biomedical Results from Skylab*, NASA, Washington, DC, 1977, pp. 235–241.
32. Leach, C.S., C.P. Alfrey, W.N. Suki, J.I. Leonard, P.C. Rambaut, L.D. Inners, S.M. Smith, H.W. Lane, and J.M. Krauhs. Regulation of body fluid compartments during short term space flight. *J. Appl. Physiol.* 81(1): 105–116 (1996).
33. Gerzer, R., M. Heer, and C. Drummer. Body fluid metabolism at actual and simulated microgravity. *Med. Sci. Sports Exercise* 28 (10, Suppl.) S32–S35 (1996).
34. Leach, C.S. A review of the consequences of fluid and electrolyte shifts in weightlessness. *Acta Astronaut.* 6: 1123–1135 (1979).
35. Norsk, P. Role of arginine vasopressin in the regulation of extracellular fluid volume. *Med. Sci. Sports Exercise* 28 (10, Suppl.) S36–S41 (1996).
36. Fritsch, J.M., J.B. Charles, B.S. Bennett, M.M. Jones, and D.L. Eckberg. Short-duration spaceflight impairs human carotid-cardiac reflex responses. *J. Appl. Physiol.* 73: 664–671 (1992).
37. Convertino, V.A., D.F. Doerr, D.L. Eckberg, J.M. Fritsch, and J. Vernikos-Danellis. Head-down bedrest impairs vagal baroreflex responses and provokes orthostatic hypotension. *J. Appl. Physiol.* 68(4): 1458–1464 (1990).
38. Cooke, W.H., J.E. Ames IV, A.A. Crossman, T.A. Kuusela, K.U. Tahvanainen, L.B. Moon, J. Drescher, F.J. Baisch, T. Mano, B.D. Levine, C.G. Blomqvist, and D.L. Eckberg. Nine months in space: Effects on human autonomic cardiovascular regulation. *J. Appl. Physiol.* 89(3): 1039–1045 (2000).
39. Hoffer, G.W., and R.L. Johnson. Apollo flight crew cardiovascular evaluations. In: *Biomedical Results of Apollo*, R.S. Johnston and L.F. Dietlein (eds). National Aeronautics and Space Administration, Washington, DC, 1977, pp. 366–371.
40. Levine, B.D., et al. *Circulation* 96: 517–525 (1997).
41. Charles, J.B., M.W. Bungo, and G.W. Fortner. Cardiopulmonary function. In *Space Physiology and Medicine*, 3rd ed. A. Nicogossian, C. Huntoon, and S. Pool (eds). Lea & Febiger, Philadelphia, 1994, pp. 286–304.
42. Charles, J.B. NASA Johnson Space Center, personal communication, 1998.
43. Grindeland, R. Bion 11 Project Scientist, NASA Ames Research Center, personal communication, June 1997.
44. Workshop to develop critical path roadmap. Johnson Space Center Space and Life Sciences Program Office and National Space Biomedical Research Institute. Held at the Center for Advanced Space Studies, Houston, Texas, January 6–8, 1998.
45. Rosenbaum, D.S., L.E. Jackson, J.M. Smith, H. Garan, J.M. Ruskin, and R.J. Cohen. Electrical alternans and vulnerability to ventricular arrhythmias. *New Engl. J. Med.* 330: 235–241 (1994).
46. Hohnloser, S.H., T. Klingenhoben, L. Yi-Gang, M. Zabel, J. Peetermans, and R.J. Cohen. T wave alternans as a predictor of recurrent ventricular tachyarrhythmias in ICD recipients: Prospective comparison with conventional risk markers. *J. Cardiovasc. Electrophysiol.* 9(12): 1258–1268 (1998).
47. Gold, M.R., D.M. Bloomfield, K.P. Anderson, D.J. Wilber, N. El-Sherif, N.M. Estes III, W.J. Groh, E. Kaufman, and R.J. Cohen. A comparison of T-wave alternans, signal averaged electrocardiography and electrophysiology study to predict arrhythmia vulnerability. *J. Am. Coll. Cardiol.* 33(Supplement A): Abstract 145A (1999).
48. Ramsdell, C.D., T.J. Mullen, G.H. Sundby, S. Rostoft, N. Sheynberg, M. Maa, G.H. Williams, and R.J. Cohen. Simulated weightlessness induces reversible changes in electrical conduction properties of the human myocardium, submitted for publication.

49. Bungo, M.W., J.B. Charles, and P.C. Johnson. Cardiovascular deconditioning during space flight and the use of saline as a countermeasure to orthostatic intolerance. *Aviation Space Environ. Med.* 56: 985–990 (1985).
50. Yelle, J. Director Cardiovascular Laboratory, NASA Johnson Space Center, personal communication, 1999.
51. Ramsdell, C.D., T.J. Mullen, G.H. Sundby, S. Rostoft, N. Sheynberg, N. Aljuru, M. Maa, R. Mukkamala, D. Sherman, K. Toska, J. Yelle, D. Bloomfield, G.H. Williams, and R.J. Cohen. Midodrine prevents orthostatic intolerance associated with simulated spaceflight. *J. Appl. Physiol.* 90: 2245–2248 (2001).
52. McTavish, D., and K.L. Goa. Midodrine: A review of its pharmacological properties and therapeutic use in orthostatic hypotension and secondary hypotensive disorders. *Drugs* 38: 757–777 (1989).
53. Ward, C.R., J.C. Gray, J.J. Gilroy, and R.A. Kenny. Midodrine: A role in the management of neurocardiogenic syncope. *Heart* 79: 45–49 (1998).
54. Wright, R.A., H.C. Kaufman, R. Perera, T.L. Opfer-Gehrking, M.A. McEllogott, K.N. Sheng, and P.A. Low. A double-blind, dose-response study of midodrine in neurogenic orthostatic hypotension. *Neurology* 52: 120–124 (1998).

CRAIG D. RAMSDELL

Dept. of Anesthesiology and Perioperative Medicine
Beaumont Hospital, Royal Oak, Michigan

RICHARD J. COHEN

Division of Health Sciences and Technology
Harvard University—Massachusetts Institute of Technology
Cambridge, Massachusetts

CHANDRA X-RAY OBSERVATORY

Introduction

The Chandra X-ray Observatory (originally called the Advanced X-ray Astrophysics Facility—AXAF) is the X-ray component of NASA's "Great Observatory" Program. Chandra is a NASA facility that provides scientific data to the international astronomical community in response to scientific proposals for using it. The Observatory is the product of the efforts of many organizations in the United States and Europe. The Great Observatories also include the Hubble Space Telescope for space-based observations of astronomical objects primarily in the visible portion of the electromagnetic spectrum, the now defunct Compton Gamma-Ray Observatory that was designed to observe gamma-ray emission from astronomical objects, and the soon-to-be-launched Space Infrared Telescope Facility (SIRTF). The Chandra X-ray Observatory (hereafter CXO) is sensitive to X rays in the energy range from below 0.1 to above 10.0 keV, corresponding to wavelengths from 12 to 0.12 nanometers. The relationships among the various

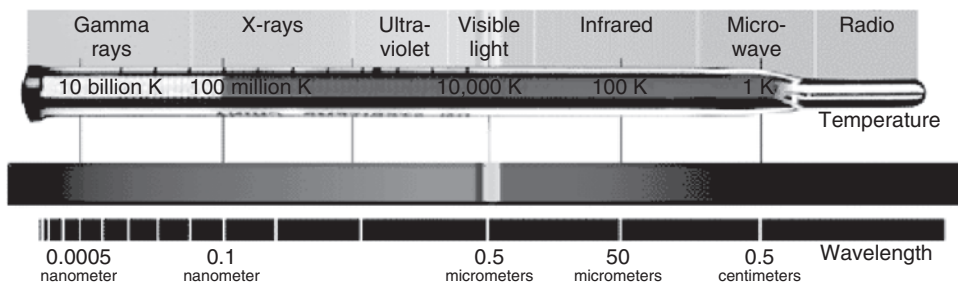


Figure 1. The electromagnetic spectrum as a function of temperature and wavelength. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

parts of the electromagnetic spectrum, sorted by characteristic temperature and the corresponding wavelength, are illustrated in Fig. 1.

In 1895, the German physicist, Wilhelm Roentgen, discovered what he thought was a new form of radiation. He called it X radiation to summarize its properties. The radiation could pass through many materials that easily absorb visible light and could free electrons from atoms. We now know that X rays are nothing more than light (electromagnetic radiation) but at high quantum energies.

Light has been given many names: radio waves, microwaves, infrared, visible, ultraviolet, X-rays, and gamma radiation are all different forms. Radio waves are composed of low-energy particles of light (photons). Optical photons—the only photons perceived by the human eye—are a million times more energetic than the typical radio photon, whereas the energies of X-ray photons range from hundreds to thousands of times higher than those of optical photons. Very low temperature systems (hundreds of degrees below 0°C) produce low-energy radio and microwave photons, whereas cool bodies like our own (about 30°C) produce infrared radiation. Very high temperatures (millions of degrees Celsius) are one way of producing X-rays.

X-ray astronomy is an extremely important field because it has been found that all categories of astronomical objects (or a subset thereof), from comets to quasars, emit X-rays. Thus learning how and why these objects produce X-rays are fundamental to our understanding of the way astronomical systems work. This, together with the large amounts of energy required to produce X-rays, makes their study interesting and exciting. A second reason that the field is so important is that the vast bulk of the matter in the Universe that we can directly observe through the electromagnetic radiation that it emits is in the very hot (temperatures of millions of degrees) X-ray emitting gas that fills the space between galaxies in clusters of galaxies (Fig. 2), the largest collections of matter in the Universe. The CXO is the prime method of gaining new information about the X-ray emission seen in the Universe.

The field of X-ray astronomy began with the discovery in 1948 that the Sun is a source of X rays. The experiment performed by scientists at the Naval Research Laboratories used photographic film mounted in a V-2 sounding rocket. Earth's atmosphere absorbs X rays, and so experiments must be performed above it. The first X-rays from a source other than the Sun were detected in 1962 by R. Giacconi and B. Rossi using a Geiger counter in a sounding rocket. For this,

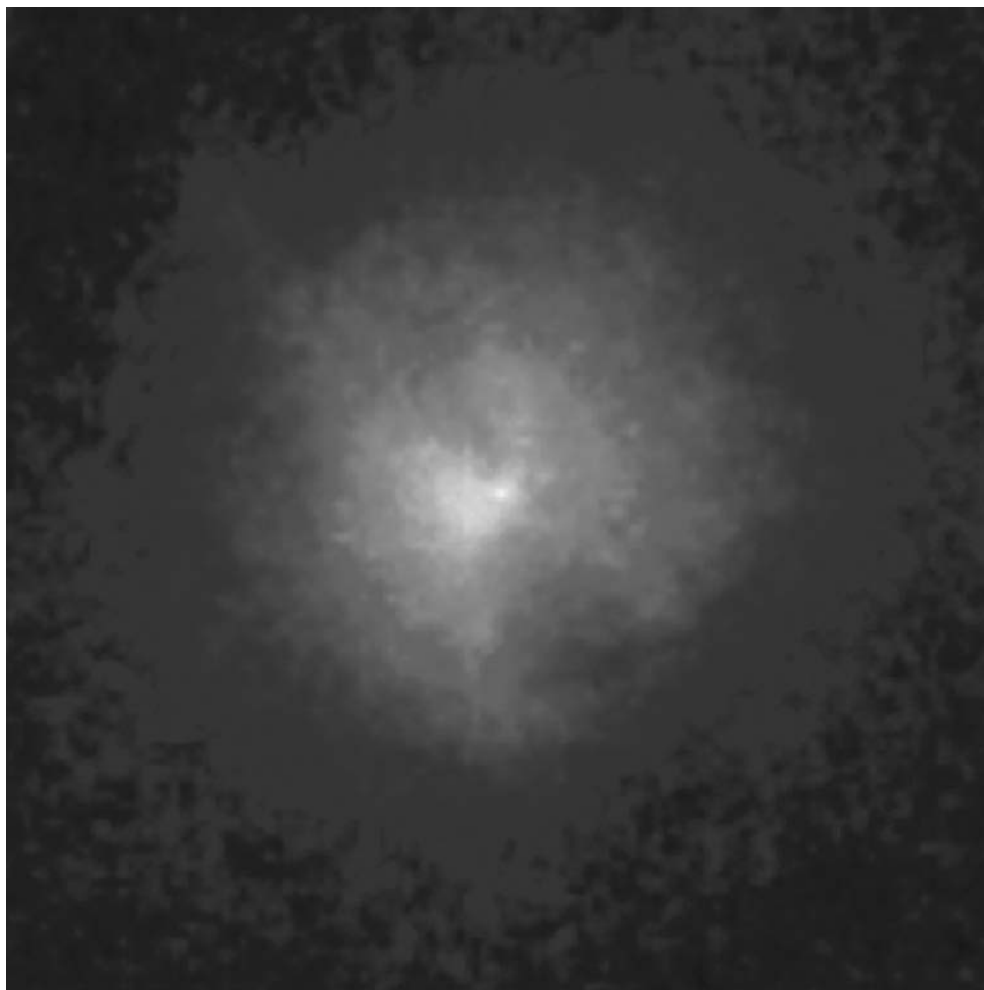


Figure 2. The Chandra X-ray image of Hydra A, a galaxy cluster 840 million light years from Earth, shows strands (blue/pink) of 35–40 million degree Celsius gas embedded in a large cloud of equally hot gas (blue) that is several million light years across. A bright white wedge of hot multimillion degree Celsius gas is seen pushing into the heart of the cluster. (Courtesy of: NASA/Chandra X-ray Center/Smithsonian Astrophysical Observatory.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

and subsequent pioneering efforts including fostering the development of the Chandra X-ray Observatory, Giacconi was awarded half of the Nobel prize for physics in 2002.

Most of the early X-ray astronomy experiments used gas-filled X-ray detectors, either Geiger or proportional counters. Locating the X-ray sources in the sky used mechanical collimation to restrict the field of view to at best about 0.5° (the apparent size of the Sun and the Moon). This crude accuracy made identification of the X-ray sources with objects seen in either the visible or the radio difficult. One notable exception in this early era was an experiment by H. Friedman and colleagues at the Naval Research Laboratories that used lunar



Figure 3. Subrahmanyan Chandrasekhar. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

occultation to show that the nebula associated with the X-ray pulsar in the Crab Nebula was extended.

X-ray astronomy began to make huge strides due to the development of focusing X-ray optics that could concentrate large amounts of X rays into small areas and form an image. The first X-ray astronomy satellite fully devoted to extrasolar X-ray imaging was NASA's second High Energy Astronomy Observatory, it was referred to as the Einstein Observatory in the scientific community, and it was launched in 1979. The prototype of the Chandra X-ray Observatory, the Einstein observatory had the first large-area grazing incidence X-ray optics and featured an angular resolution of about 10 seconds of arc, although only a small fraction of the X rays reflected by the telescope was within this central core due to scattering caused by the roughness of the reflecting surfaces.

The next major advance in X-ray focusing was the German/USA/UK mission Rosat (1990–1998; see, e.g., Truemper, 1983). It was equipped with an imaging X-ray telescope of about 4 seconds of arc angular resolution and three interchangeable imaging X-ray detectors, and it made numerous contributions.

The CXO was named in honor of the late Indian-American Nobel laureate, Subrahmanyan Chandrasekhar (Fig. 3), nicknamed “Chandra” which means “moon” or “luminous” in Sanskrit. He was one of the foremost astrophysicists of the twentieth century and, in 1983, was awarded the Nobel prize for his studies of the physical processes important to the structure and evolution of stars.

The Observatory and Its Instrumentation

An artist's drawing of the CXO is shown in Fig. 4. The CXO has three major parts, as shown in Fig. 5: (1) the X-ray telescope or High-Resolution Mirror Assembly



Figure 4. Artists rendering of the CXO. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

(HRMA), whose mirrors focus X rays from celestial objects; (2) the science instruments—the advanced CCD imaging camera (ACIS) and the high-resolution camera (HRC) that record the X rays and two sets of objective transmission gratings (OTG) discussed below; and (3) the spacecraft, which provides functions such as power and telemetry necessary for the telescope and the instruments to work.

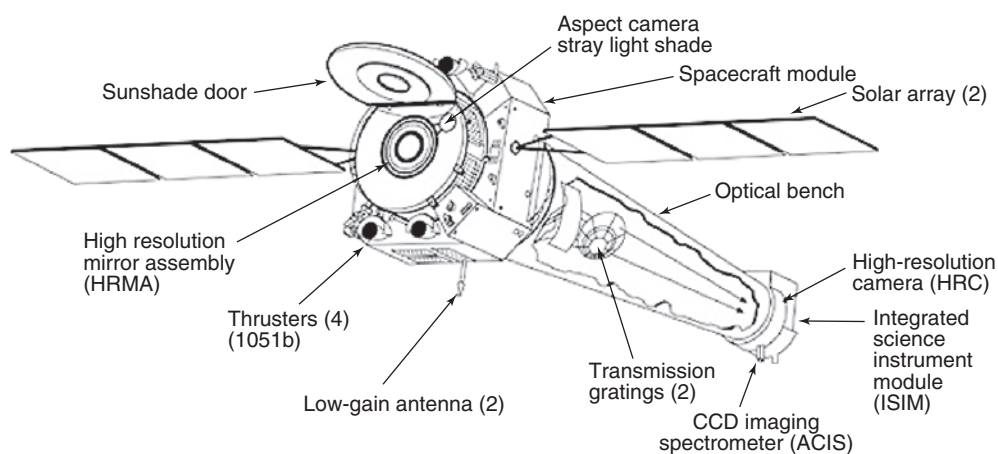


Figure 5. Line drawing of the major components of the CXO. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The building and operation of this 1.5-billion-dollar facility has been a marvel of modern technology and ingenuity. Overall program management and technical and scientific oversight are provided by NASA's Marshall Space Flight Center. The Marshall Center was ably assisted by the Smithsonian Astrophysical Observatory (SAO). The prime contractor was the company, TRW, Inc., which was responsible for the spacecraft construction and integrating of all subsystems into the Observatory. Major subcontractors and their principal functions were Raytheon Optical Systems—telescope grinding and polishing; Optical Coating Laboratories, Inc.—telescope coating; Eastman Kodak Corporation—telescope assembly and alignment; Ball Aerospace and Technology Corp.—science instrument accommodation module and aspect system.

The spacecraft systems for the CXO are fairly standard; they have modest (32,000 bits/second) data rates and moderate pointing accuracy and stability. These latter two were feasible by employing an onboard, visible-light-sensitive, aspect camera which is used on the ground after the data are taken to correct the position of the X rays detected to sky coordinates. This is possible because the X-ray cameras on CXO record the position of each X-ray detected.

Because Earth's atmosphere absorbs X rays, the CXO was placed high above it. This meant that the ultraprecise mirrors and detectors, together with the sophisticated electronics that conveys the information back to Earth, had to withstand the rigors of a rocket launch and operate in the hostile environment of space.

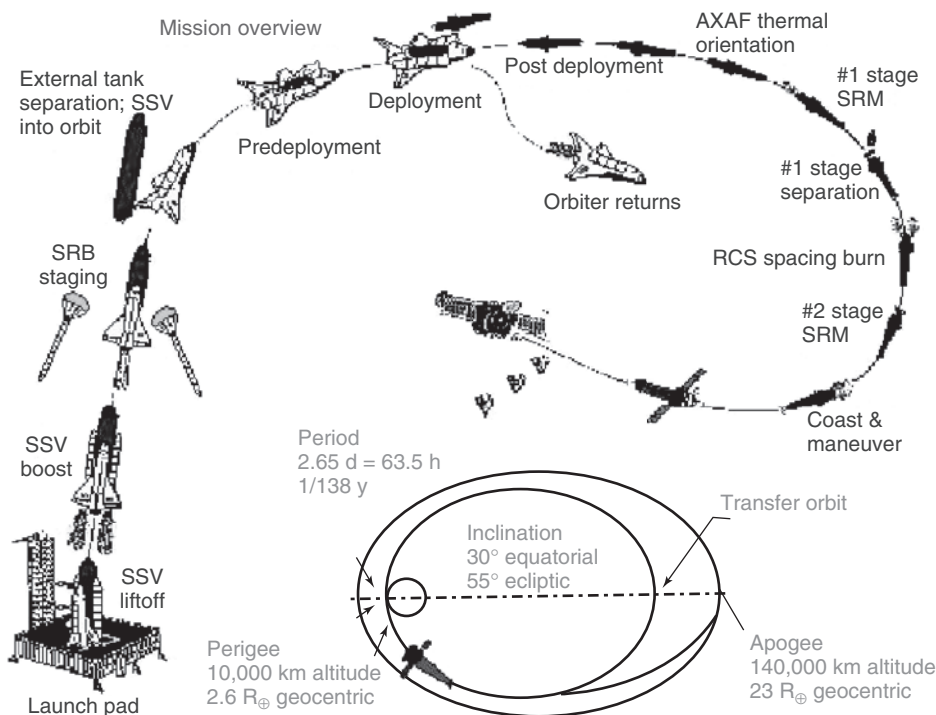


Figure 6. CXO launch sequence and orbit. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Chandra's unusual orbit, shown in Fig. 6, was achieved after initial deployment by the Space Shuttle Columbia, Eileen Collins (Fig. 7) commanding. This particular Shuttle launch was especially noteworthy because Commander Collins was the first female commander. Initial deployment was followed by a boost into a high Earth orbit by an inertial upper stage built by the Boeing Corporation. Final placement into the orbit used a built-in propulsion system. The orbit, which has the shape of an ellipse, takes the spacecraft more than a third of the way to the moon before returning to its closest approach to Earth, 10,000 kilometers (6,214 miles). The time to complete an orbit is about 65 hours. The spacecraft spends about 75% of its orbit above the belts of charged particles that surround Earth. Uninterrupted observations as long as 55 hours are possible.

Because of their high quantum energy, X rays do not easily reflect from mirrors. However, reflection can take place when the angle of incidence is shallow. This property can be exploited to build optical systems that can bring X rays to a common focus. A particular design, that used for the CXO, is illustrated in Fig. 8. The design, referred to as a Wolter type 1, uses a paraboloid of revolution followed by a hyperboloid of revolution—two reflections are necessary to bring objects away from the axis of symmetry into focus. Nesting several of these



Figure 7. Commander Eileen Collins. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

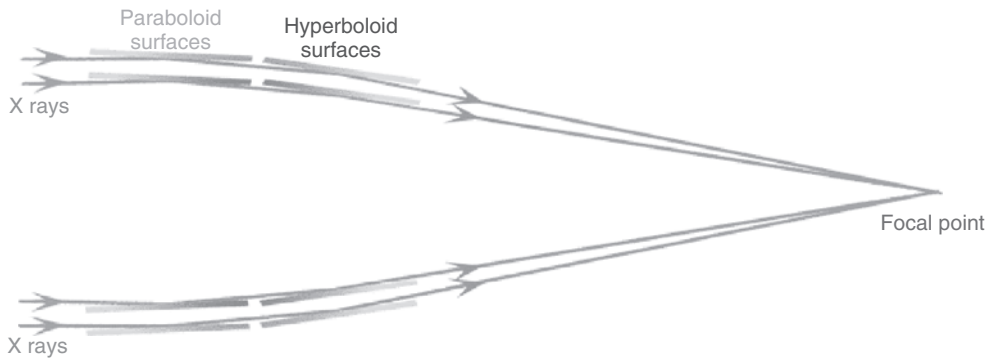


Figure 8. Illustration of a cross section of two nested, Wolter I, X-ray optics. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

pairs of paraboloids and hyperboloids increases efficiency for collecting X rays. The CXO has four such systems whose diameters range from 0.65 m (2.13 ft) to 1.2 m (3.94 ft). Each optical element is 0.84 m (2.75 ft) long. The elements are constructed of Zerodur, a glassy ceramic, and together weigh more than 2000 pounds.

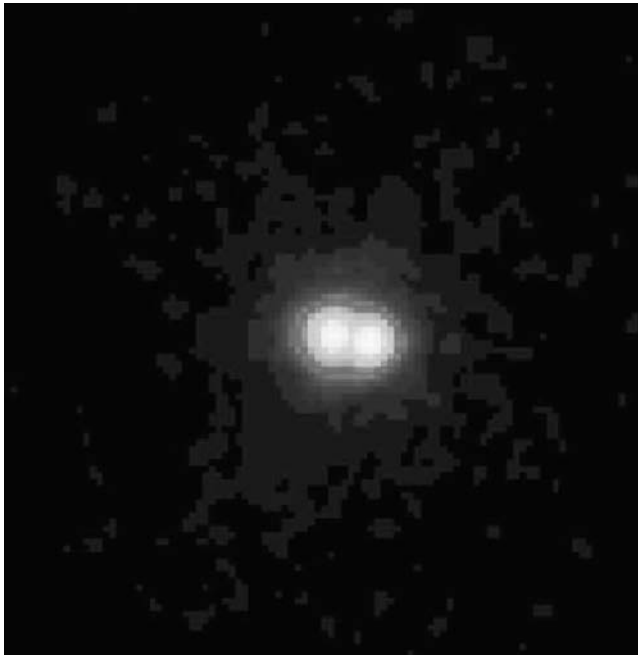


Figure 9. Chandra's image of what had been a puzzling X-ray source in the globular star cluster M15 shows that it is not a single system but two sources that are so close together (2.7 seconds of arc) that they were undistinguished by previous X-ray telescopes (courtesy of NASA/White and Angelini, 2001). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The most unique attributes of the CXO X-ray optics, those that makes this observatory so powerful, are the angular resolution (the ability to distinguish two separate objects very close to each other) and the high efficiency by which photons are collected within the narrow spot that defines its ability to resolve different objects. This high efficiency results from the smoothness of the X ray reflecting surfaces so that the X-rays are not scattered from the intended path. The angular resolution is better than 0.5 seconds of arc, that is, the letters of an X-ray-emitting STOP sign can be distinguished at a distance of 12 miles! The surface roughness is measured in angstroms. If the state of Colorado were as smooth as Chandra's mirrors, Pike's Peak would be less than 1 inch tall. The angular resolution of Chandra is significantly better than any previous, current, or even currently planned X-ray observatory. Figures 9 and 10 and illustrate the utility of the superb angular resolution and low scattering.

The science instruments aboard the Observatory and the organizations that led their development are the advanced CCD imaging spectrometer (ACIS)—Penn State University and the Massachusetts Institute of Technology (MIT); the high-resolution camera (HRC)—Smithsonian Astrophysical Observatory (SAO); the high-energy transmission grating (HETG)—MIT; and the low-energy transmission grating (LETG)—Space Research Institute of the Netherlands and the Max Planck Institute for Extraterrestrial Physics in Garching, Germany.

Either (but not both simultaneously) of the two transmission gratings can be commanded in place directly behind the telescope. When in position, X rays are

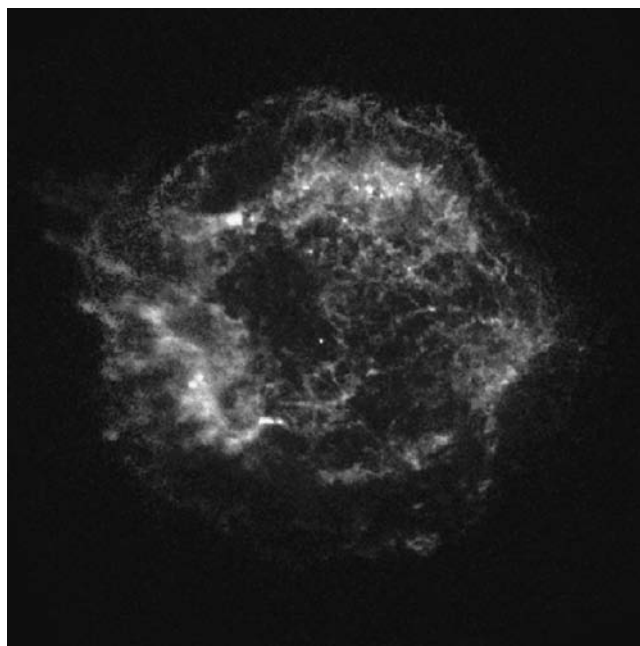


Figure 10. Chandra image of the supernova remnant Cassiopeia-A based on about three-quarters of an hour of data. The point source at the center, previously undetected, simply leaps out of the Chandra image (courtesy of NASA/Chandra X-ray Center). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

diverted (dispersed) according to their energy, along one dimension, from their normal path. The X-ray image now represents, to a high degree of accuracy, the energy of the incident photons. Determining the energy is called spectroscopy.

The ACIS detectors record X-ray images and can, to a certain degree, also determine the energy of the incident X-ray but not as well as the gratings. The spectroscopic advantage of the ACIS detectors is that they are far more efficient for detecting X rays when the gratings are not in place. There are two ACIS detectors; either can be placed at the focus of the telescope by command. The ACIS-I is made of a 2×2 array of front-illuminated (FI), 2.5-cm-square, X-ray sensitive, charge-coupled devices (CCDs) analogous to those found in visible-light-sensitive digital cameras. ACIS-I provides high-resolution spectrometric imaging across a 17-arcmin-square field of view. ACIS-S, a 6×1 array of 4 FI CCDs and two back-illuminated CCDs, is mounted along the dispersion direction of the transmission gratings and serves both as the primary read-out detector for the HETG, and, using one of the back-illuminated CCDs that can be placed at the aim point of the telescope, also provides high-resolution spectrometric imaging extending to lower energies, but across a smaller (8-arcmin-square) field than the ACIS-I.

As in the ACIS, there are two HRC detectors that may be moved into place to record X-ray images. Both are microchannel-plate detectors that consist of millions of tiny tubes of cesium-iodide-coated glass. These detectors record the position of the X rays, but unlike the ACIS, can barely distinguish energies. The HRC-I array is a 10-cm square plate that has a field of view of 31 arcmin square. Comprising three rectangular segments (3×10 cm each) and mounted end-to-end along the dispersion direction, the HRC-S serves as the primary read-out detector for the LETG. A important advantage of the HRC for certain scientific experiments is its ability to determine the time when the X-ray event was detected to high a much higher precision (microseconds) than achievable with the ACIS (milliseconds).

Scientific Results

X rays are produced by highly energetic processes—thermal processes in plasmas at temperatures of millions of degrees—or non thermal processes, such as synchrotron emission (realized when charge particles are accelerated by magnetic fields) or scattering of visible light or radio waves from very hot electrons. Consequently, X-ray sources are frequently exotic: supernova explosions and remnants, whose explosion shocks the ambient interstellar medium or a pulsar, a rotating neutron star, powers the emission; disks of accreting nearby material or jets around stellar-mass neutron stars or black holes; accretion disks or jets around massive black holes in the nucleus of galaxies; hot gas in galaxies and in clusters of galaxies that traces the gravitational field and can be used to determine the mass. These are all examples of sites of, and methods for, producing X rays. Here we give several examples of observations by the CXO that illustrate its capability for investigating these processes and objects.

Chandra's capability for high-resolution imaging enables studies of the structure of extended X-ray sources, including supernova remnants (see Fig. 10),

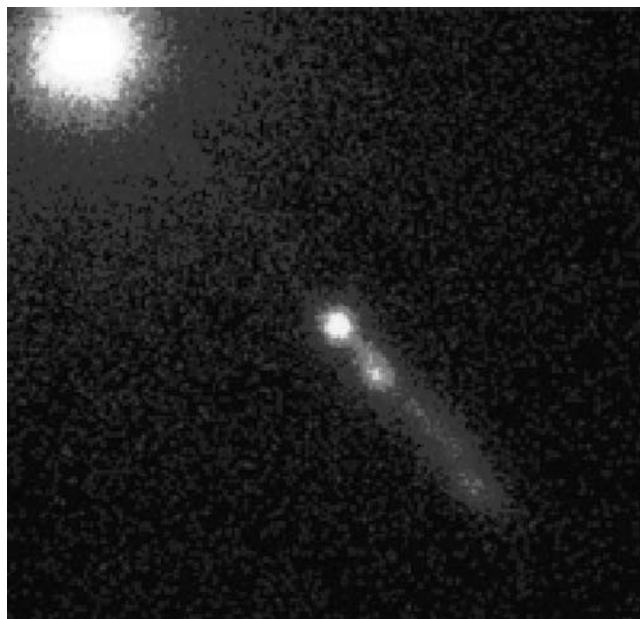


Figure 11. This Chandra image, 24 arc seconds on a side, shows details in the powerful jet shooting from the quasar 3C273, providing an X-ray view into the area between 3C273's core and the beginning of the jet (courtesy NASA/Chandra X-ray Center/Smithsonian Astrophysical Observatory/Marshall et al., 2001). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

astrophysical jets (Fig. 11), and hot gas in galaxies and clusters of galaxies (Fig. 2). The capability for spectrometric imaging allows studies of structure in X-ray intensity and also in temperature and chemical composition. Through observations by Chandra, scientists have begun to address several of the most exciting topics in contemporary astrophysics.

Comets and Planets. Besides the Sun, which is too bright to be observed safely by the Observatory, other known X-ray sources in our solar system include Earth, the Moon, Venus, Jupiter and some of its moons, and comets. The X-rays from the Moon, Venus and, to a certain extent, Earth are due to the fluorescence of solar X rays that strike the atmosphere. The discovery image of X-ray emission from Venus (Fig. 12) by Chandra shows a half crescent due to the relative orientation of the Sun, Earth, and Venus. Solar X-rays are absorbed in the Venusian atmosphere above the surface of the planet that knock electrons out of the inner parts of atoms and excite the atoms to a higher energy level. The atoms almost immediately return to their lower energy state and emit a characteristic or “fluorescent” X-ray. Similarly, ultraviolet light produces the visible light of fluorescent lamps. The origin of the X-ray emission from Jupiter (Fig. 13) is more complex and not completely understood. Important ingredients are the presence of a magnetic field and rotation of the planet. Studies using Chandra should help provide a better understanding of the physical processes involved.

Normal Stars. The CXO has optical instrumentation aboard that provides an “aspect solution,” which shows where the observatory was pointing during an

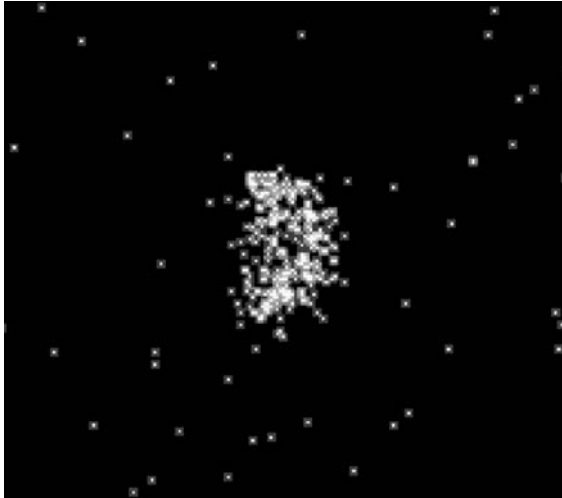


Figure 12. Chandra image, the first X-ray image ever of Venus, shows a half crescent due to the relative orientation of the Sun, Earth, and the planet—the Sun is to the right. X-rays from Venus are produced by fluorescent radiation from oxygen and other atoms in the atmosphere about 130 kilometers above the surface (courtesy of NASA/Max Planck Institute for Extraterrestrial Physics/Dennerl et al., 2002). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

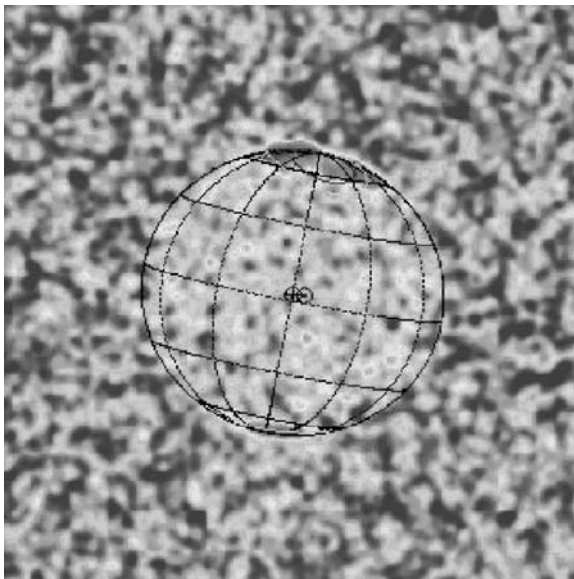


Figure 13. Chandra image of X-ray emission from Jupiter. The red colors indicate the highest intensity. The bulk of the emission takes place near the magnetic poles (courtesy of NASA/Marshall Space Flight Center/Gladstone et al., 2002). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

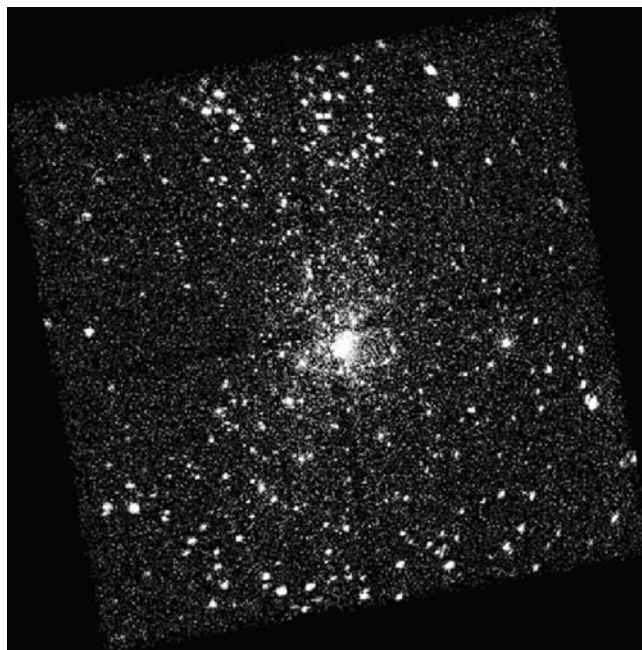


Figure 14. The CXO image of the Orion star cluster. The region shown in this image is about 10 light years across. The bright stars in the center are part of the Trapezium, an association of very young stars whose ages are less than a million years (courtesy of NASA/Pennsylvania State University). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

observation to an accuracy of about 1 second of arc. However, it is usual for Chandra observations to detect the X-ray emission from several stars that have much more accurate cataloged positions based on their optical emission. Using the stellar positions as a local reference, the remaining X-ray sources in the field can be positioned far more precisely to about 0.2 to 0.4 seconds of arc. One example of using such positions is the determination of the positions of the more than 1000 X-ray sources in the Orion Nebula star cluster (Fig. 14). It is interesting that this single Chandra image contains more X-ray sources than had been detected in the first 15 years of X-ray astronomy. The Orion Trapezium, a region only 3 light years across at the core of the Orion Nebula star cluster, contains very young stars and therefore offers astronomers a view into a region where stars are born. The cluster is composed of stars whose a median age is only around 300,000 years, and, at a distance of 1400 light years, is one of the nearest star-forming regions. The CXO was used to identify X-ray emission from individual stars and found that almost all of these young stars were much hotter than expected.

The Center of Our Galaxy. Precise positioning was critical for the unique identification of the X-ray emission from the object known as Sagittarius A*, the source at the extremely crowded center of our own Milky Way galaxy (Fig. 15). This source is a moderately small black hole of about 3 million solar masses, and is very faint compared to other galactic nuclei. X rays are emitted just beyond

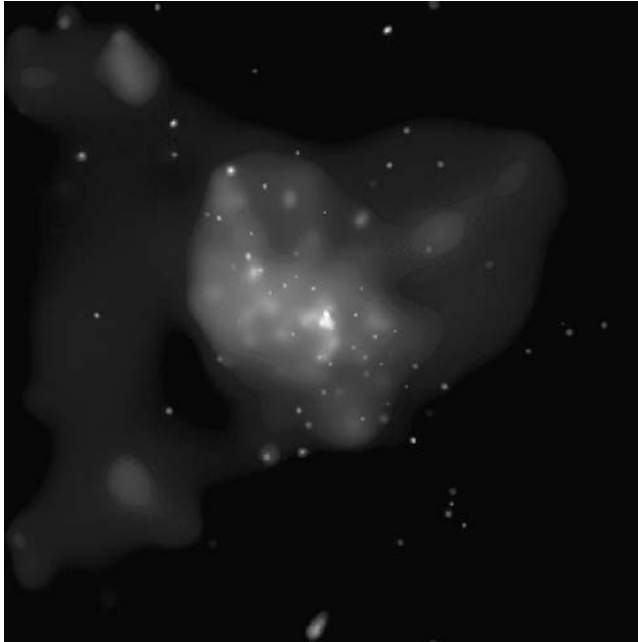


Figure 15. The ACIS-I image of the galactic center. This image shows the central region of our Milky Way Galaxy as seen by the CXO. The bright, point-like source at the center of the image was produced by a huge X-ray flare that occurred in the vicinity of the black hole at the center of our galaxy (courtesy of NASA/Massachusetts Institute of Technology/Baganoff et al., 2001). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the horizon within which no light can escape, and thus the black hole may be “seen.” A bright flare was detected by the CXO on 27 October 2000, whose X-ray intensity increased by more than a factor of 10 for about 3 hours and then rapidly dipped on a timescale of 10 minutes. The rapid variation in X-ray intensity indicates that the flare was due to material as close to the black hole as Earth is to the Sun. This is compelling evidence that matter falling toward the central black hole is fueling energetic activity at the center of our galaxy.

Supernova Remnants. Another example of Chandra’s ability to provide high-contrast images is exemplified by the now classic image of the remains of the remnant formed by the implosion of a star, a supernova, called the Crab Nebula and its compact core, a rotating neutron star that pulses. The image (Fig. 16) shows the intricate structure produced by the pulsar interacting with the local environment. This image is not static; observations by the CXO spaced out over time have shown the dynamic motions that take place as a result of the interaction of the pulsar at the center of the remnant and the local environment.

Globular Clusters. One of the most striking examples of the power of Chandra X-ray imaging is in the spectacular Chandra images of globular clusters. Figure 17 is an ACIS-I image of the globular cluster 47 Tucanae. The left panel is composed of red/green/blue images derived from X rays recorded in different energy bands and shows the central portion of the cluster. The

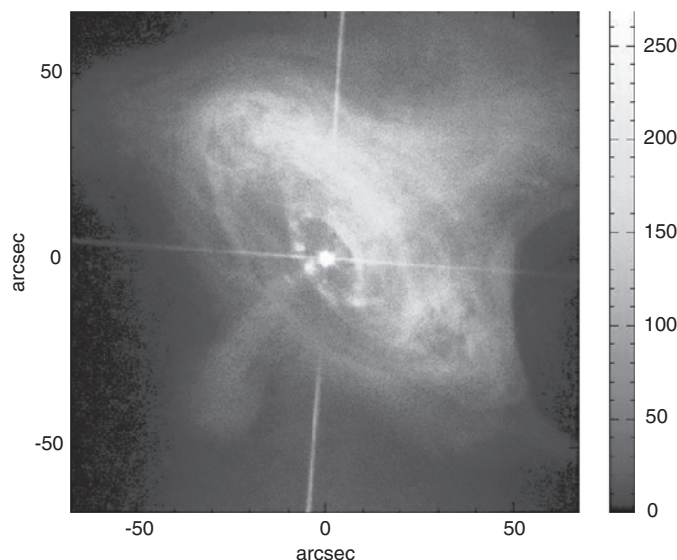


Figure 16. Chandra image of the Crab pulsar and nebula made using the LETGS and the HRC-S. The nearly horizontal line in the figure is a cross-dispersed spectrum of the pulsar produced by the LETG fine support bars. The nearly vertical line is the LETG-dispersed spectrum from the pulsar (courtesy of NASA/Marshall Space Flight Center/Weisskopf et al., 2000). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

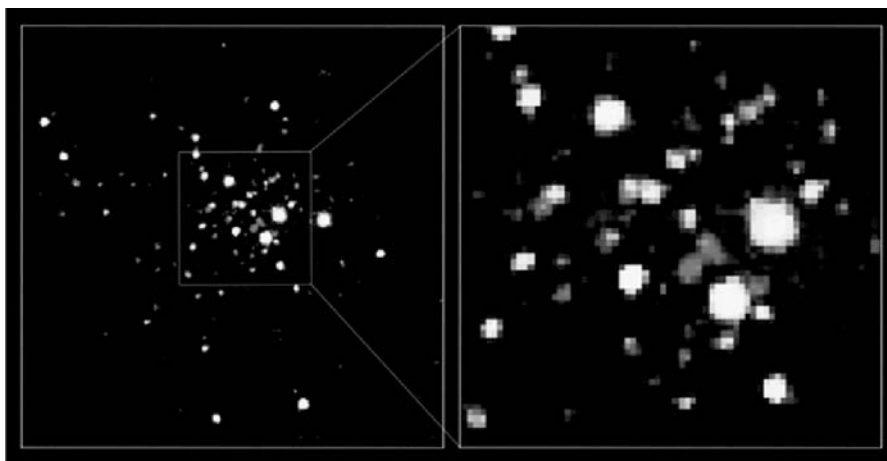


Figure 17. Chandra image of 47 Tucanae. The left image covers the central $2' \times 2.5'$. The further enlargement of the central region to the right is $35'' \times 35''$. The different colors represent the dominant X-ray energy range: low-energy X-ray emission (red), intermediate-energy X-ray emission (green), and high-energy X-ray emission (blue). The white sources are bright in each of these ranges. The faint red sources are mostly millisecond pulsars, whereas the bright white sources are mostly binaries containing white dwarfs that produce X-rays by pulling matter from normal stars (courtesy of NASA/CfA (Center for Astrophysics)/Grindlay et al., 2001). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

enlargement at the right is the central core. Some 108 sources, excluding the central core, are detected. This number is more than ten times that found by previous X-ray satellites.

These Chandra images of 47 Tucanae provided the first complete census of the brighter X-ray producing stars in the core of the globular cluster. As the oldest stellar systems in galaxies, globular clusters are laboratories for stellar evolution. Nearly all objects in the Chandra images are “binary systems,” in which a normal, Sun-like star orbits a collapsed star, either a white dwarf or neutron star from which X-rays are emitted. The data also reveal the presence of many “millisecond pulsars,” neutron stars that rotate extremely rapidly between 100 to nearly 1000 times a second.

The image, associated spectra, and measured time variability reveal more about the binary content and stellar, as well as dynamic, evolution of a globular cluster than achieved by all previous X-ray observations of globular clusters combined.

Normal Galaxies. In addition to mapping the structure of extended sources and the diffuse (extended) emission due to the presence of hot gas in galaxies, the high angular resolution of the CXO permits studies of ensembles of discrete sources, which would otherwise be impossible owing to the large number of sources crowded together in a small region of the sky (source confusion). A beautiful example comes from observations of the center of the Andromeda galaxy (M31) shown in Fig. 18. The image shows what used to be considered the emission associated with the black hole at the center that is now resolved into five distinct sources. A most interesting consequence is that the emission from the region surrounding the central black hole is unexpectedly faint relative to the mass, as in the Milky Way.

M81 (NGC 3031) is a spiral galaxy at a distance of approximately 3.6 Mpc. The galaxy was observed by using the ACIS-S instrument on Chandra (Fig. 19). Previously, the galaxy had been observed using both the Einstein and Rosat X-ray observatories, and about 30 sources had been detected. This Chandra observation detected 97 X-ray sources, 41 in the bulge of the galaxy and 56 sources in the disk. Twenty-one of these latter sources were close to, or in, the spiral arms. The sheer number of X-ray sources detected by Chandra in a typical nearby galaxy makes studies of the global properties of X-ray source populations possible.

In many galaxies, observations using Chandra have detected one or more sources which are so bright that if truly associated with the galaxy and if the emission is not highly beamed, then the object must have a mass substantially greater than that of a neutron star and is thus likely to be a black hole whose mass is more than 25 times the mass of the Sun. Such “ultraluminous” sources appear to be quite common in nearby galaxies, and there are too many of examples of such objects to give much credence to the idea that they are all more local than one might think. If they are local (closer), then they are not nearly so powerful as they appear and would not be considered “ultraluminous.” Because of the ubiquity, the concept of an intermediate-mass black hole is gaining some acceptance. Lower mass black holes are born in explosions of massive stars (supernova); very high mass (hundreds of thousands of solar masses or larger) black-holes are born in (form in) the nuclei of galaxies. Precisely how an “intermediate-mass” black hole is created is not completely understood.

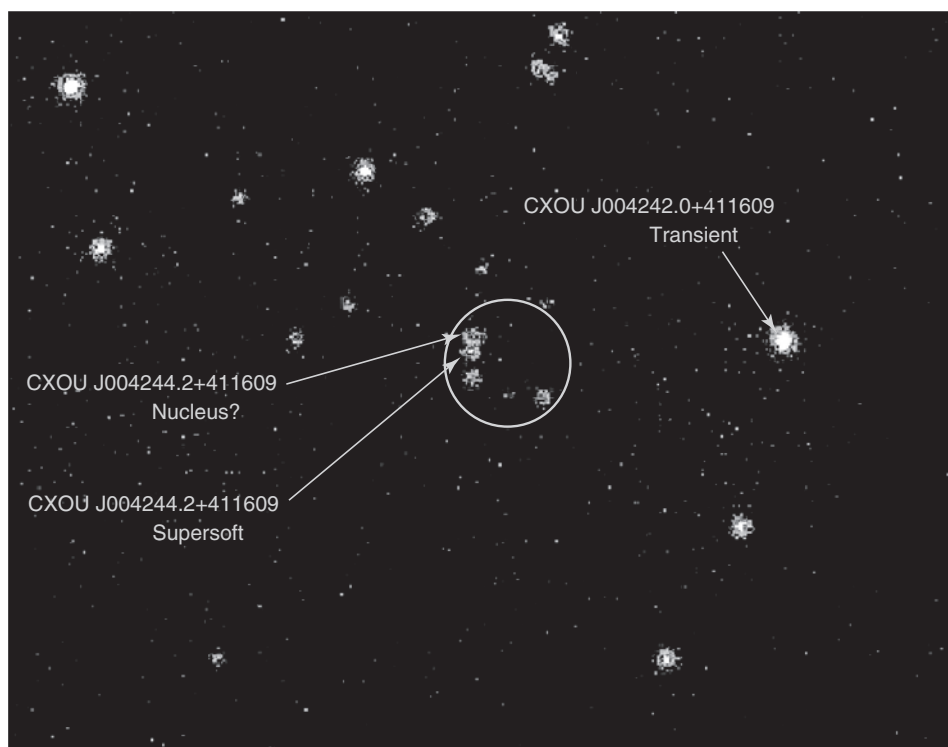


Figure 18. Central region of M31 observed with the ACIS-I. The circle is 5 arcseconds in radius and illustrates the previous location of the X-ray nucleus. The old X-ray nucleus is resolved into five individual sources, and the source labeled CXO J004244.2 + 411609 is within 0.15 arcseconds of the black hole at the center of the galaxy (courtesy NASA/Smithsonian Astrophysical Observatory/Chandra X-ray Center/Garcia et al., 2002). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Quasars. Quasars look like normal stars through an optical telescope. Not until the 1950s, when radio astronomy was first developed, did astronomers discover that these objects were well outside our own galaxy and were therefore emitting massive amounts of radio energy. These objects are called “quasars,” short for “quasi-stellar radio sources.” Quasars are among the most distant, energetic objects ever observed. Individual quasars are brighter than hundreds of galaxies put together, yet many are smaller than the size of our own solar system. Radio astronomers use a system of numbers to name objects in the sky. 3C273 was named in the 3rd Cambridge catalog as the 273rd radio source identified. 3C273, along with 3C48, were the first quasars to be identified as such. These objects had bizarre optical spectra unlike any ever studied before. In 1963, astronomers Maarten Schmidt (3C273) and Jesse Greenstein and Thomas Matthews (3C48) noticed that these spectra made sense only if the objects were simply moving away from us, thus causing the apparent wavelengths to shift to the longer (redder) wavelengths (hence redshift)—in the case of 3C273 at about one-tenth the speed of light. Quasars are now considered a subset of active galaxies (called active galactic nuclei or AGN) powered by the presence of a massive black holes at their centers.

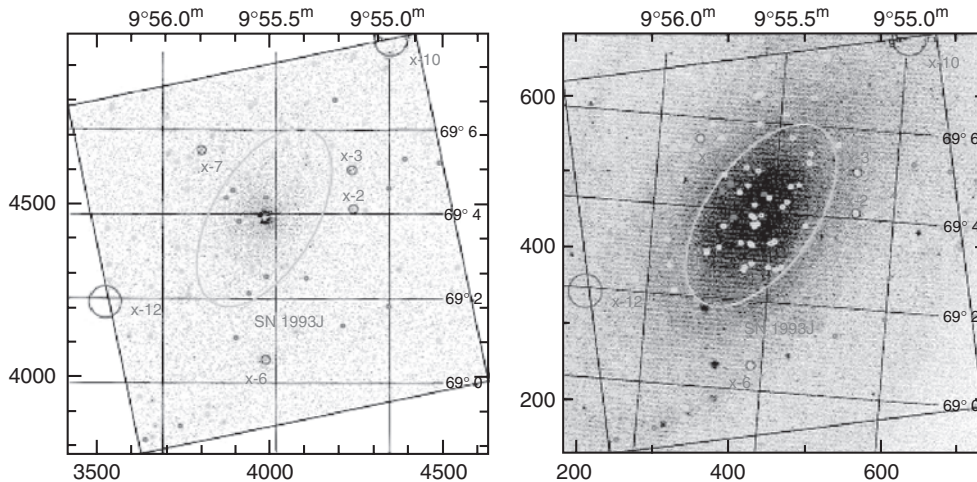


Figure 19. Chandra observations of M81. Left: X-ray image with contours. Right: X-ray contours on optical image (courtesy of NASA/Marshall Space Flight Center/Tennant et al., 2001). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

More detailed and modern studies of quasars show that there are high-powered jets associated with these objects and often the material in the jets moves at velocities very close to the speed of light. Exactly how this happens is a topic of major astrophysical interest. Furthermore, instead of seeing a smooth stream of material driven from the core of the quasar, most optical, radio, and X-ray observations have revealed inconsistent, “lumpy” clouds of gas. The Chandra image (Fig. 11) of 3C273, however, shows, for the first time, a continuous X-ray flow emanating from the core. One would like to learn how and why matter is so violently ejected from near the quasar’s core. The energy emitted probably comes from gas that falls toward the supermassive black hole at the center of the quasar but is channeled by strong electromagnetic fields. The black hole itself cannot be observed directly, but scientists hope to discern properties of the black hole by studying the jet.

Clusters of Galaxies. Chandra observations of clusters of galaxies frequently show us structures that have characteristic angular scales of a few arcseconds. Before Chandra, it was thought that the X-ray emission from galaxy clusters arose from a fairly simple system. The Hydra A radio galaxy (3C 218), shown in Fig. 2, was observed early in the Chandra mission. This very early image shows large areas of low X-ray brightness, indicating the presence of regions of low density or cavities in the hot gas. Similar, but even more dramatic cavities were found in the Perseus cluster shown in Fig. 20.

The X-ray Background—the Chandra Deep Surveys

The first sounding rocket flight in 1962 that detected the brightest X-ray source in the sky, other than the Sun, also detected a general background of X radiation (Giacconi et al., 1962). Since this first experiment, the detailed nature of the

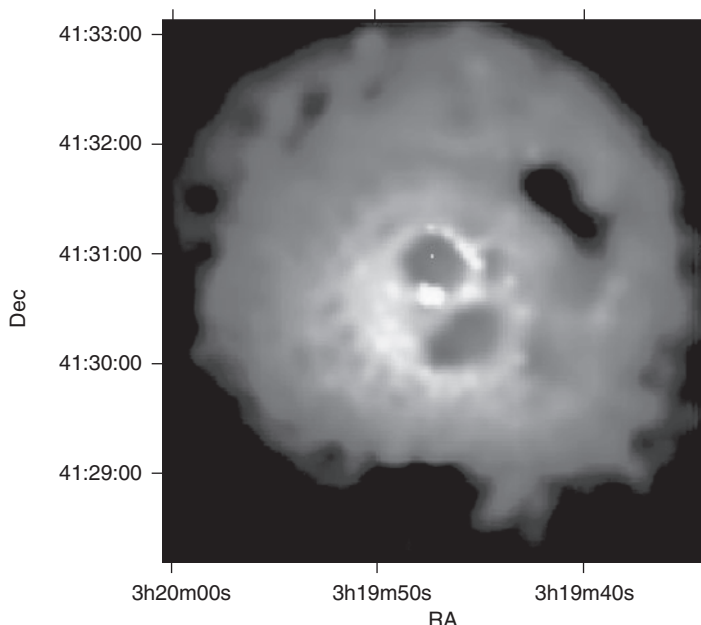


Figure 20. Chandra image of the X-ray core of the Perseus cluster of galaxies (courtesy of NASA/Institute of Astronomy/Fabian et al., 2000). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

background radiation had been a puzzle. The lack of any distortion of the energy spectrum of the much cooler cosmic microwave background could be used to eliminate the possibility that the X-ray background glow was truly diffuse (Mather et al., 1990). Thus, the glow must have consisted of sources so faint as to be unresolved by the telescopes of the time. The question remained—what faint sources were contributing to this background glow? Observations by previous X-ray satellites such as ROSAT were a major step in resolving a significant fraction (about 75%) of the glow at low energies into discrete objects (Hasinger et al., 1998) and found that the sources reside mainly in active galaxies at redshifts from 0.1 to 3.5. The Japanese ASCA satellite observations extended the search for sources in the 2–10 keV band and resolved about 30% of the background glow (Ueda et al., 1998).

Chandra is ideally suited to search for faint sources. Two very deep exposures, one of one million, the other of two million seconds have been accomplished to date. These exposures are referred to as the Chandra Deep Fields. A portion of the image of one, the Chandra Deep Field North, is shown in Fig. 21. About 350 sources were detected in each of the one-million-second surveys. The most distant source detected (so far) is a quasar at a redshift of 5.2. The majority of the X-ray sources beyond a redshift of 0.5 are active galaxies that have massive black holes at their centers. The Chandra deep surveys extend the study of the background to signals levels more than an order of magnitude below that which could be previously achieved and have confirmed that most of this background is due to a variety of different types of discrete sources. At very low

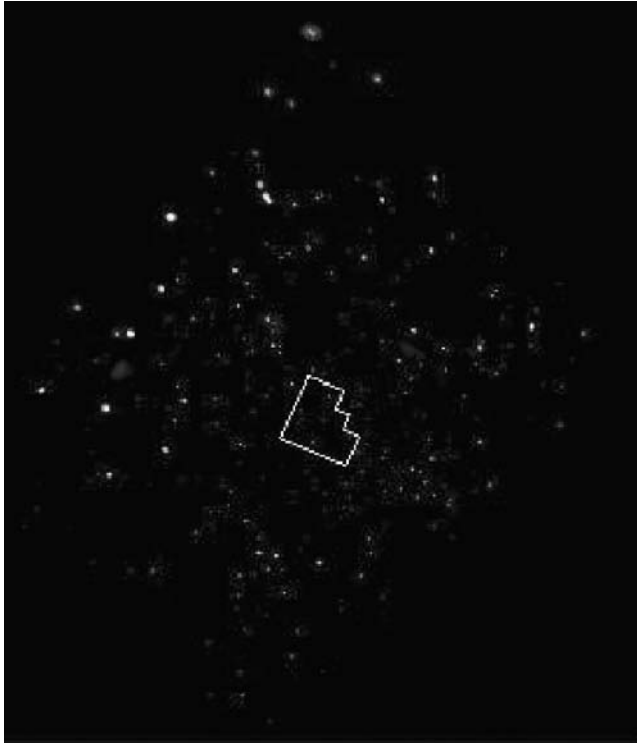


Figure 21. Chandra “true -color” ACIS image of the Chandra Deep Field North. This image has been constructed from the 0.5–2.0 keV band (red) and 2.0–8.0 (blue) images. Part of the Chandra Deep Field North has also been surveyed in visible light using the Hubble Space Telescope, and this region is shown in outline (courtesy of NASA/Pennsylvania State University/Hornschemeier et al., 2001). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

energies, there is somewhat diffuse component that results from warm-hot intervening matter (referred to as WHIM), concentrated regions of matter that formed in the early stages of the Universe when the first galaxies and galactic clusters formed.

The spectrum of the X-ray background had also been considered puzzling because it was found that the majority of the bright AGN have an energy spectrum different from that of the background itself. One of the purposes of the Chandra Deep Surveys has been to explore the spectra of the sources, because the faint sources must modify the spectrum from the average of the bright AGN sample. The faint source spectrum was estimated by adding the spectra of individual sources detected in the surveys. This has now been accomplished and the paradox resolved (see, e.g., Rosati et al., 2002).

Sources of Gamma-Ray Bursts. Cosmic gamma-ray bursts (GRBs) were discovered in the late 1960s by satellites designed to detect gamma rays produced by atomic bomb tests on Earth. The bursts appear as a brilliant flash of gamma rays that lasts seconds to minutes. The bursts are often, but not always, followed by afterglows observable at X-ray (always), optical, and radio



Figure 22. Chandra image of the X-ray afterglow from gamma-ray burst GRB010222 (courtesy of NASA/Center for Nuclear Research/Garmire et al.). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

wavelengths (sometimes). Using one of the experiments on the Compton Gamma-Ray Observatory, astronomers took an important step in understanding the sources of these bursts by determining that the objects that produce them are not in our own galaxy. This discovery implied that the sources of these bursts are extremely energetic. The source of this tremendous energy is unknown. Models currently in vogue involve a fireball of energy where matter moves at nearly the speed of light. The source of this rapidly moving matter is unknown; theories include the merging of neutron stars, or black holes, or the collapse of an extremely massive star.

The CXO is being used to help to solve the mystery of gamma-ray bursts by studying X-ray afterglow (Fig. 22). An example of this is the Chandra observation of the gamma-ray burst that took place on 22 February 2001. The Chandra data supported the model in which a very massive star has exploded, a “hypernova.”

High-Resolution Spectroscopy. Owing to their unprecedented clarity, Chandra images are visually striking and provide new insights into the nature of X-ray sources. Equally important are Chandra’s unique contributions to high-resolution dispersive spectroscopy. High-resolution X-ray spectroscopy is the essential tool for diagnosing conditions in hot plasmas. It provides information for determining the temperature, density, elemental abundance, and ionization stage of X-ray emitting plasma. The excellent spectral resolution of the Chandra gratings isolates individual spectral lines (i.e., features at specific energies), which would overlap at lower resolution. The spectral resolution also enables the

determination of flow and turbulent velocities, by measuring Doppler shifts, shifts in the apparent energy, and the widths of the lines produced by these motions. Spectroscopy using the CXO gratings achieves its best energy resolution for nonextended (point) sources. Thus, observations that use the Chandra grating have concentrated on, but are not limited to, X-ray emission from the hot regions surrounding stars (stellar coronae), X-ray sources in binary systems, and active galactic nuclei.

Web Site

The official CXO site on the World Wide Web is <http://www.chandra.harvard.edu>. This site has numerous Chandra images, as well as information about the Observatory and X-ray astronomy.

BIBLIOGRAPHY

1. Baganoff, F.K., M.W. Bautz, W.N. Brandt, G. Chartas, E.D. Feigelson, G.P. Garmire, Y. Maeda, M. Morris, G.R. Ricker, L.K. Townsley, and F. Walter. *Nature* 413: 45 (2001).
2. Dennerl, K., V. Burwitz, J. Englhauser, C. Lisse, and S. Wolk. *Astron. Astrophys.* 386: 319 (2002).
3. Fabian, A.C., J.S. Sanders, S. Etti, G.B. Taylor, S.W. Allen, C.S. Crawford, K. Iwasawa, R.M. Johnstone, and P.M. Ogle. *MNRAS*, 318: L65 (2000).
4. Garcia, M.R., et al. *Astrophys. J.* in press.
5. Garmire, G.P., A.B. Garmire, L. Piro, E. Schlegel. *GCN* 1000: 1 (2001).
6. Giacconi, R., et al. *Phys. Rev. Lett.* 9: 439 (1962).
7. Giacconi, R. The Einstein X-ray Observatory. *Sci. Am.* 242: 80 (1980).
8. Gladstone, G.R., J.H. Waite, D. Grodent, W.S. Lewis, F.J. Crary, R.F. Elsner, M.C. Weisskopf, T. Majeed, J.-M. Jahn, A. Bhardwaj, J.T. Clarke, D.T. Young, M.K. Dougherty, S.A. Espinosa, and T.E. Cravens. *Nature* 415: 1000 (2002).
9. Grindlay, J.E., C. Heinke, P.D. Edmonds, and S.S. Murray. *Science* 290: 2292 (2001).
10. Hasinger, G., R. Burg, R. Giacconi, M. Schmidt, J. Trumper, and G. Zamorani. *Astron. Astrophys.* 329: 482 (1998).
11. Hornschemeier, A.E., W.N. Brandt, G.P. Garmire, D.P. Schneider, A.J. Barger, P.S. Broos, L.L. Cowie, L.K. Townsley, M.W. Bautz, D.N. Burrows, G. Chartas, E.D. Feigelson, R.E. Griffiths, D. Lumb, J.A. Nousek, L.W. Ramsey, and W.L.W. Sargent. *Astrophys. J.* 554: 742 (2001).
12. Marshall, H.L., D.E. Harris, J.P. Grimes, J.J. Drake, A. Fruscione, M. Juda, R.P. Kraft, S. Mathur, S.S. Murray, P.M. Ogle, D.O. Pease, D.A. Schwartz, A.L. Siemiginowska, S.D. Vrilek, and B.J. Wargelin. *Astrophys. J.* 549: L167 (2001).
13. Mather, J.C., E.S. Cheng, R.E. Eplee, Jr., R.B. Isaacman, S.S. Meyer, R.A. Shafer, R. Weiss, E.L. Wright, C.L. Bennett, N.W. Boggess, E. Dwek, S. Gulkis, M.G. Hauser, M. Janssen, T. Kelsall, P.M. Lubin, S.H. Moseley, Jr., T.L. Murdock, R.F. Silverberg, G.F. Smoot, and D.T. Wilkinson. *Astrophys. J.* 354: L4 (1990).
14. Rosati, P., P. Tozzi, R. Giacconi, R. Gilli, G. Hasinger, L. Kewley, V. Mainieri, M. Nonino, C. Norman, G. Szokoly, J.X. Wang, A. Zirm, J. Bergeron, S. Borgani, R. Gilmozzi, N. Grogan, A. Koekemoer, E. Schreier, and W. Zheng. *Astrophys. J.* 566: 667 (2002).

15. Tennant, A.F., K. Wu, K.K. Ghosh, J.J. Kolodziejczak, and D.A. Swartz. *Astrophys. J.* 549: L43 (2001)
16. Truemper, J. *Adv. Space Res.* 2: 241 (1983).
17. Ueda, Y., T. Takahashi, H. Inoue, T. Tsuru, M. Sakano, K. Ohta, Y. Ishisaki, Y. Ogasaka, K. Makishima, and T. Yamada. *Astronomische Nachrichten* 319: 4 (1998).
18. Weisskopf, M.C., J.J. Hester, A.F. Tennant, R.F. Elsner, N.S. Schulz, H.L. Marshall, M. Karovska, J.S. Nichols, D.A. Swartz, J.J. Kolodziejczak, and S.L. O'Dell. *Astrophys. J.* 536: L81 (2000).
19. White, N.E. and L. Angelini. *Astrophys. J.* 561: L101 (2001).

MARTIN C. WEISSKOPF
NASA Marshall Spaceflight Center
Huntsville, Alabama

CIVIL LAND OBSERVATION SATELLITES

Introduction

The first, and still the most dramatic, example of seeing where we live from the vantage point of space was the image of a pale blue and white globe hanging all alone in an infinite expanse of darkness taken through the window of Apollo 8 on its way to the Moon in December 1968. Three and a half years later on 23 July 1972, NASA launched ERTS-1, later renamed Landsat-1. This was the first civil imaging satellite that had enough resolution, 80 meters, to image human-scale activities; that is everything bigger than a football field. Its more capable successors now provide anyone, anywhere, the ability to see images of motorcycle size objects any place on the globe at almost any time. As of November 2002, there are 23 satellites in orbit whose resolutions range from 30 meters down to 0.6 meter. They are being operated by the United States, France, India, Korea, Canada, China/Brazil, the European Space Agency, and three private corporations, two U.S. and one Israeli. The number of satellites and the number of their national and private sponsors show that civil land observation satellites have arrived at the point where they are now another permanent payoff of the space age. It will take much longer than the Weather, Communication and Global Positioning System (GPS) satellites for their full economic and social effects to be felt, but they are already forcing national and international discussions on the effects on nations, corporations, and individuals of the worldwide transparency that these satellites will inevitably bring (1).

The Big Picture

Figure 1 provides the orbital history of every satellite whose sensors can provide Earth images, both optical and radar, and whose resolutions are equal to and better than Landsat 1's 80 meters. It shows quite clearly that there have been three distinct periods in the history of civil land imaging satellites. For 13 years,

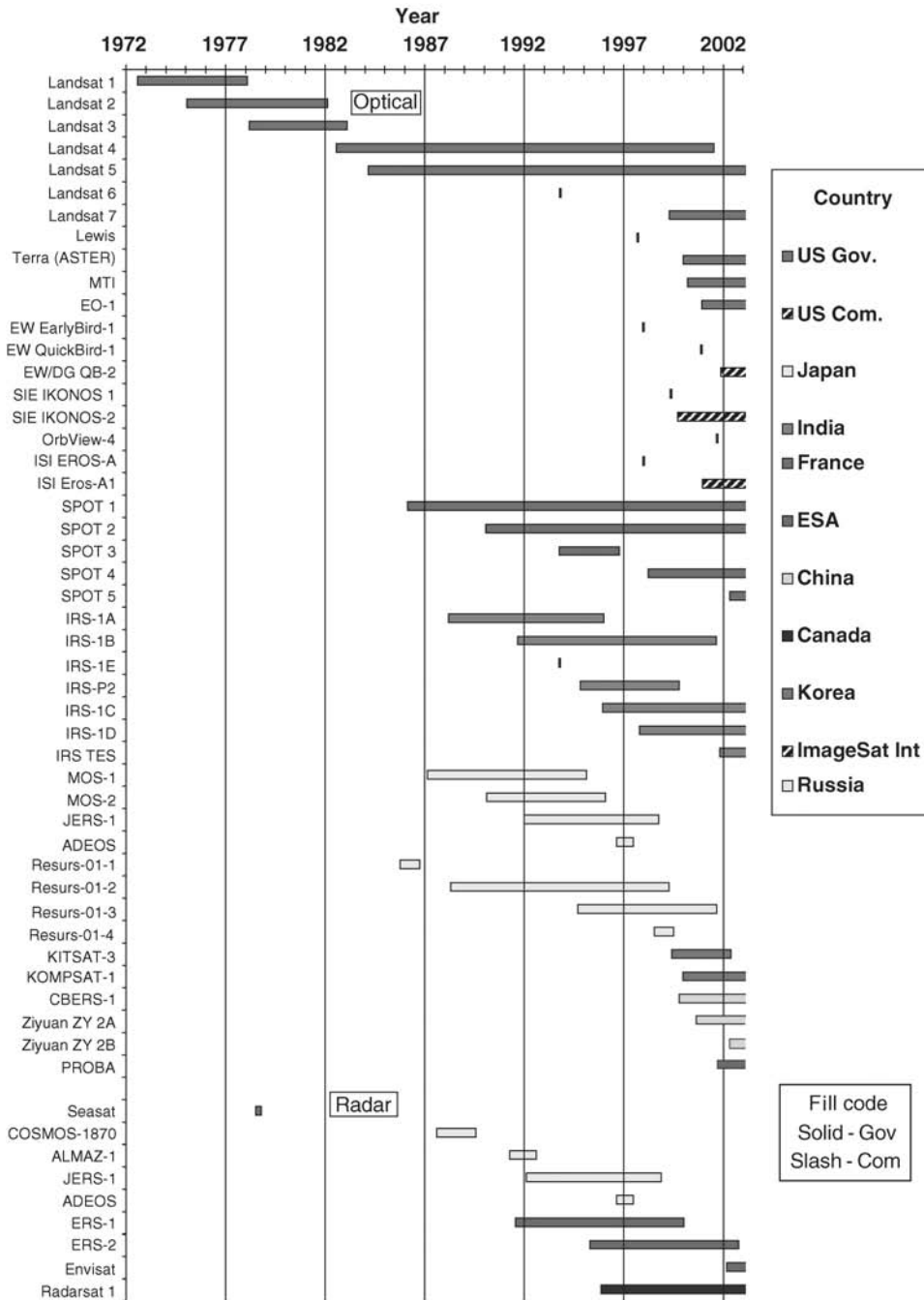


Figure 1. Orbital history of all optical and radar land imaging satellites that have Landsat or better resolution. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the United States' five optical and one radar satellites were the only such systems in orbit that provided civil data. The second period started with the launches of Russia's Resurs-01-1 in late 1985 and of France's SPOT 1 early in 1986 and extends up to the first launch attempts by American commercial systems in 1997. During this period, there were 20 foreign launches and one failed U.S. attempt. The third period begins in 1997 when high-resolution commercial systems entered, includes NASA's return to Landsat and research launches, and continues to the present day. The following discussion will provide some of the details of these periods to illuminate why they occurred as they did and will provide some sense of where the development and use of civil land imaging satellites might take us in the future. First, however, there are a few facts about satellite and sensor technology that will be helpful in understanding their development.

The Technology

This article discusses the special set of Earth orbiting satellites designed to acquire detailed images of the global land surface. It is the high-resolution subset of the larger family of Earth sensing satellites that image the land and oceans on a kilometer scale and measure the characteristics of the atmosphere to record the weather and to explore the complex interrelationships among the atmosphere, the oceans, and the land surface that cause the weather and our climate. The following describes the satellite, sensor, and data technologies involved.

Satellites. Essentially all current and planned imaging satellites operate in sun-synchronous polar orbits and cover Earth about 14 times a day, as illustrated in Fig. 2. Sun-synchronous means that the satellite crosses the equator at the same time on every orbit. This characteristic means that all images have the same Sun angle and, therefore, shadow characteristics thus eliminating one interpretation variable. (The Sun angle changes as the seasons change.) The crossing times vary between 9 and 10:30 A.M. The current satellite orbits range between 460 and 904 kilometers; the lower range is chosen by balancing the need for a high enough orbit to ensure a long satellite lifetime and a lower orbit for higher resolution. The higher limit is a balance between the desire for the broad coverage area provided by high orbits and the higher resolution provided by lowering the orbit.

Sensors. Except for one Russian intelligence film based satellite (Spin-2) used for a commercial mission, all commercial optical sensors have been and are electronic imagers; they focus the light from the ground onto the telescope's focal plane, either by sweeping it across a small line of sensors in a system called a whisk-broom scanner or by using the satellite's movement along its orbit to sweep the image across a much larger set of sensors. This latter design, called a push-broom scanner, is the dominant sensor used today with one prominent exception: Landsat 7 has retained the whisk-broom design first flown on Landsat 1.

However, from the viewpoint of the uses to which the data are to be put, the choice of scanner type is not as critical as the four basic imaging characteristics of the total satellite system, that is, spatial resolution (how small an object can be perceived), scene size (how large an area is captured in each image), temporal

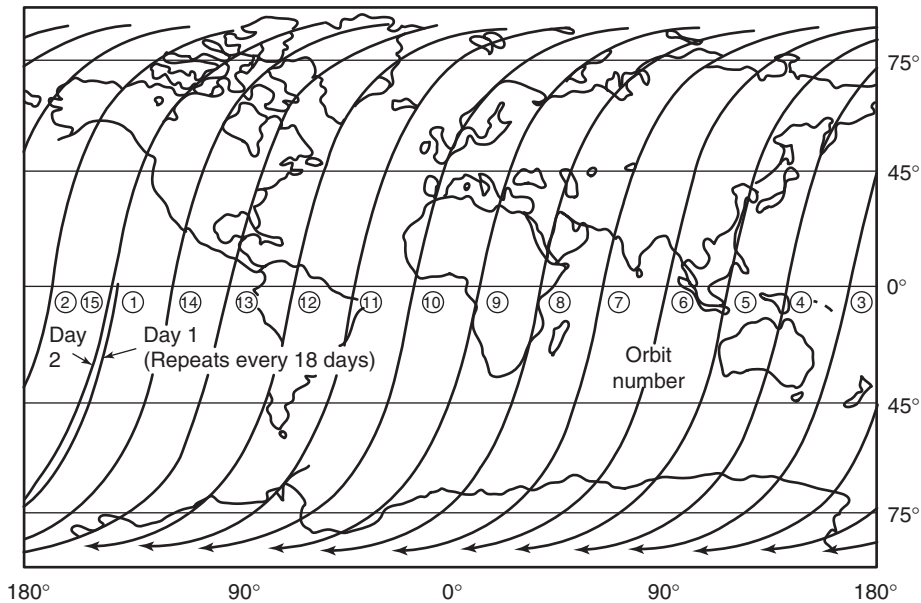


Figure 2. Typical Landsat daily ground trace.

resolution (how quickly a scene can be imaged again), and spectral resolution (the number and location of the spectrum bands measured). Each will be discussed in turn.

Spatial Resolution. The size of the imaged feature on the ground that the satellite can resolve is largely a function of the sensor's telescope design and the satellite's orbital altitude. Resolution is the most important of the three system characteristics for most users because it defines the scale and, therefore, the type of applications the system can serve. Scale is particularly important in making maps. Figure 3 provides a comparison of images of the U.S. Capitol taken at the resolutions of current satellites.

Scene Size. The width on the ground imaged by the sensor/satellite system is called its swath and serves as a surrogate for scene size since the along track image can be taken continuously and scenes are created for data management reasons by arbitrarily cutting the strips into lengths about equal to their swath width. Large scene sizes are essential for the repeat imaging of the regional and global areas that is required to study the changes in our land cover and land uses. In general, the higher the resolution, the smaller the swath.

Temporal Resolution. For fixed nadir-pointing satellites like Landsat that can only image the land beneath their orbit, the time to return to image a given area is almost entirely a function of its swath width. The TIROS weather satellite and its 1000-meter resolution AVHRR imager has a ground swath of 2500 kilometers and images the entire globe daily. Landsat has a 185-kilometer swath and repeats its global cycle every 16 days.

For satellites that can point their sensor to either side of the ground track, the return visit time is a function of both the basic orbit repeat time and the



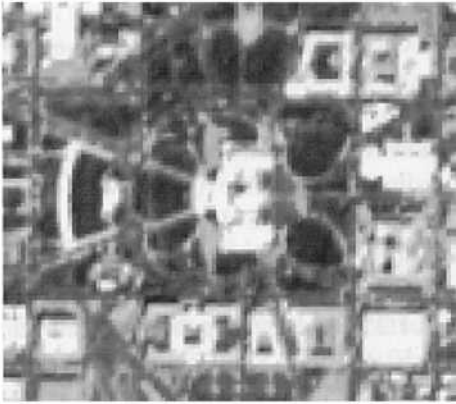
OrbView Pan - 1 meter GSD



OrbView color - 4 meter GS



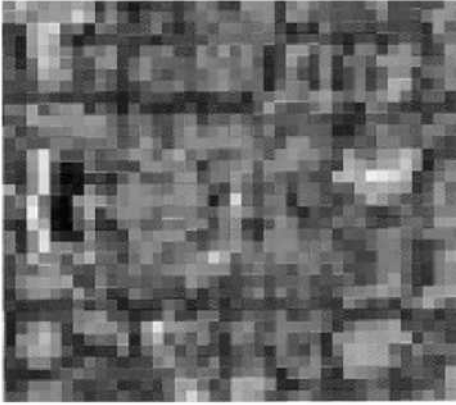
IRS-1C Pan - 6 meter GSD



SPOT Pan - 10 meter GSD



SPOT XS - 20 meter GSD



Landsat TM - 30 meter GSD

Figure 3. Comparison of image quality as a function of the resolutions of current satellites. Reproduced by the kind permission of Northrop Grumman Information Technology Inc. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

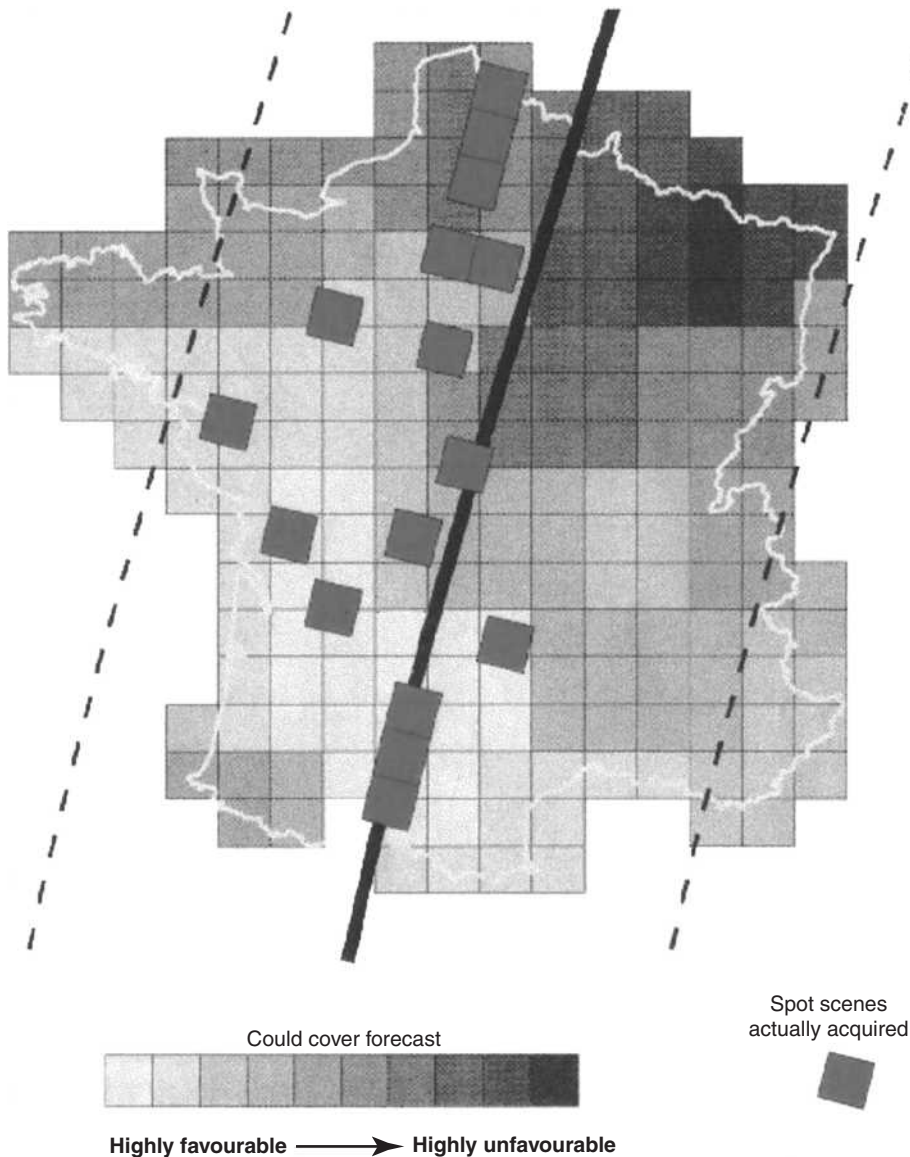


Figure 4. Example of acquisition optimization over France showing the impact of cloud cover forecasts and the satellite's mirror-pointing capabilities (each square is 60 km on the side). Reproduced by the kind permission of SPOT image. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

range of the off angle pointing. SPOT uses a mirror to point the sensor and can provide a variety of scene location patterns (see Figure 4) as well as increase its 22 day nadir image repeat time to 3 or 4 days. The smaller higher resolution satellites point their sensor by tilting the spacecraft to achieve 2 to 3 day repeat cycles. However, because these systems have very small swath widths, between 8 and 16 kilometers, they would take 6 months or more to image the globe. Please

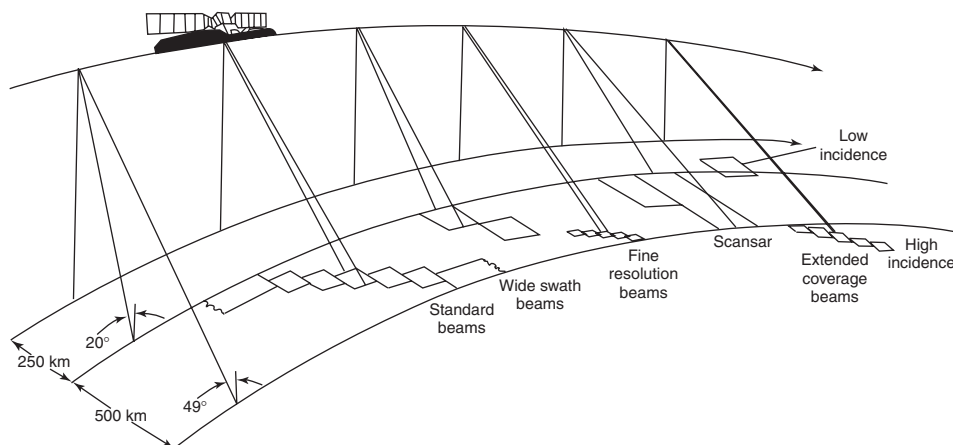


Figure 5. Imaging modes available from Radarsat.

keep in mind that the world's average 50% cloud cover means that in practice getting repeat cloud-free images could take from 2 to 3 times longer than the days quoted above.

The location, pattern, and size of radar images are controlled by electronically programming the radar's antennae and can easily be varied in a number of ways. This flexibility is illustrated in Fig. 5.

Spectral Measurement. The number and the placement of the spectral bands that are sensed are critical to the ability to classify surface features automatically, be they vegetable, mineral, or man-made. This feature of multispectral sensors has extended the use of images beyond their original mapping and object identification functions. The repeated classification of the continuing changes in forests, rangeland, and farmland at the state, country, and global scales, so important to understanding our environment, could not be accomplished by the eyes of photo interpreters alone. Figure 6 presents the number and location of the spectral bands for the major satellite classes. Figure 7 provides examples of the way Landsat band data can be combined in three band combinations to bring out various features in the scene.

The Road to Landsat

As noted before, the practical beginning of civil land imaging satellites was the 1972 launch of Landsat 1. But the process that preceded and finally resulted in that launch was long, complex, and acrimonious. It is important because its history illustrates many issues that are still not fully resolved to this day.

Humankind has probably recognized the military advantage of occupying the high ground since people first descended from the trees. As early as 1783, it was obvious to Benjamin Franklin, when, as Ambassador to France, he observed the first Montgolfier balloons in flight over Paris and predicted their use in "conveying Intelligence" about "an Enemy's Army" (2). After the Wright brothers

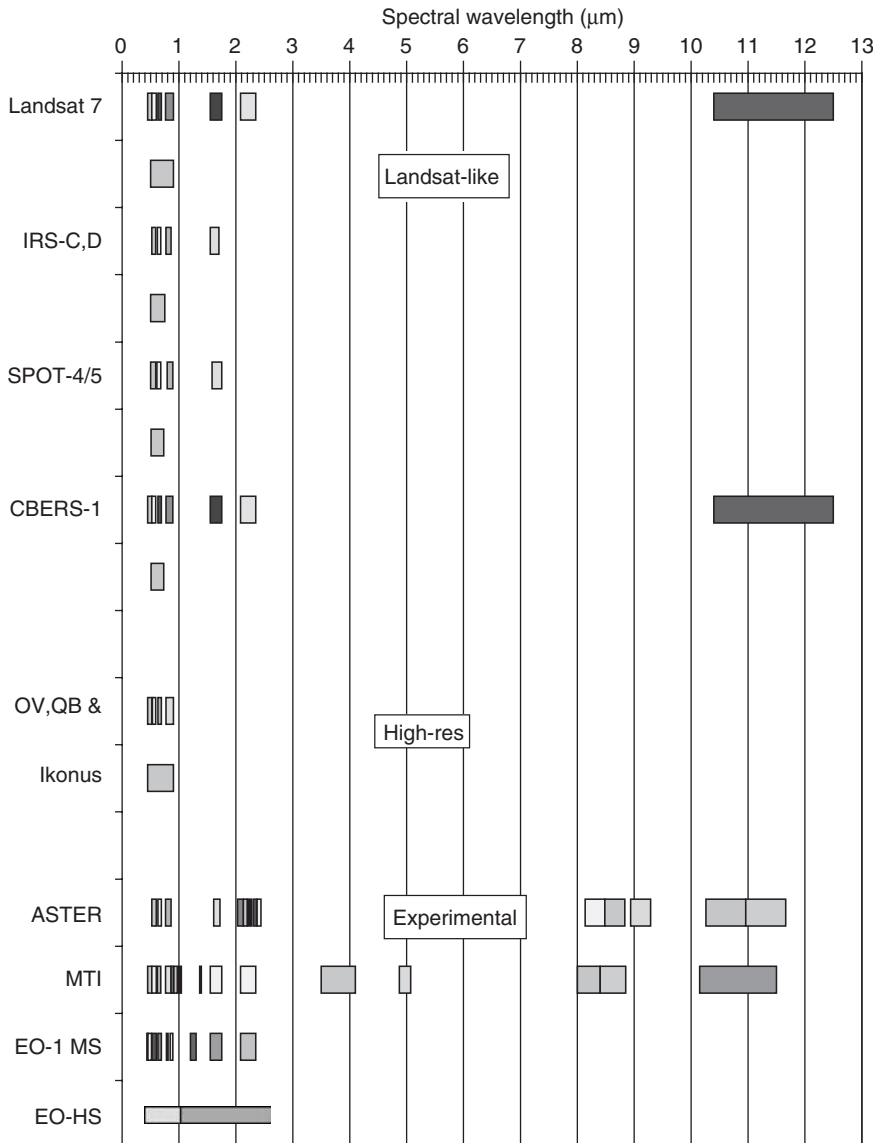


Figure 6. Number and location of the spectral bands for several satellites. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

demonstrated flight, the incredibly rapid development of the airplane in World War 1 was due almost entirely to the importance of its unique reconnaissance capability. (The development of fighter aircraft, dogfights, and aces resulted when each side tried to deny the enemy the use of air reconnaissance.) Bombers dominated aircraft use in the Second World War, but reconnaissance was still a major air function, and infrared imaging was developed out of the need to identify camouflage. The cold war caused the United States to develop the fastest and highest flying aircraft in the world for the sole purpose of photoreconnaissance.

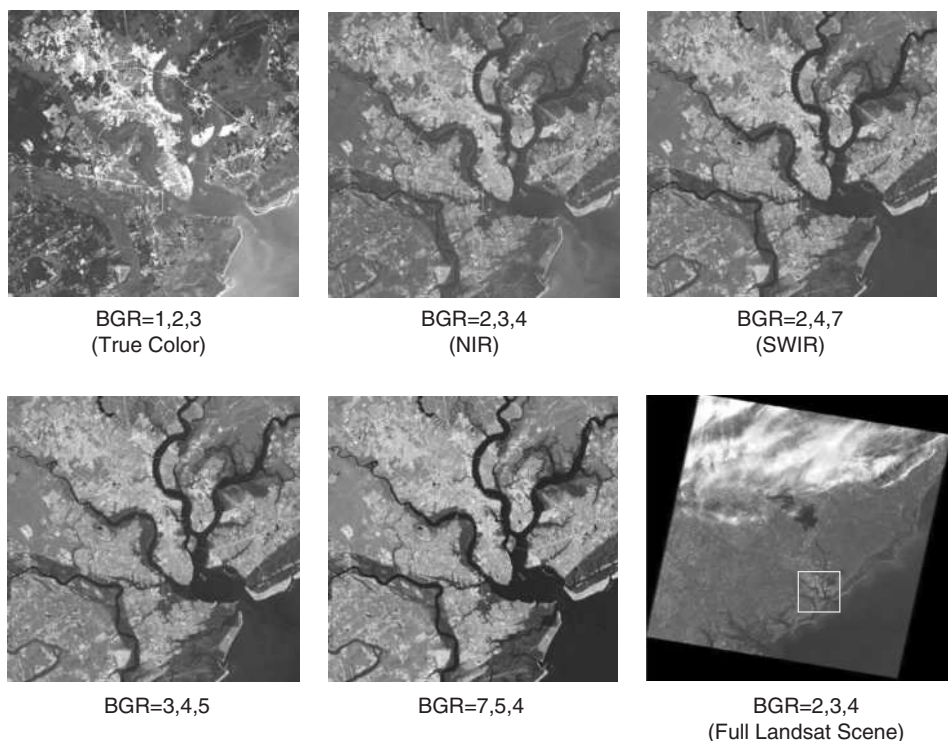


Figure 7. Color images generated by assigning different TM bands to the blue, green and red guns of the color monitor. Provided by Stacy Bunin and Mitretek Systems. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Although the Soviet launch of Sputnik in October 1957 shocked the American public, the only surprise to the science and military communities was that the Soviets did it first. In 1955, both the Soviets and the United States had announced plans to orbit an Earth satellite to support the 1957/58 International Geophysical Year (IGY). The American announcement mentioned the beginning of the ill-fated Vanguard Program that President Eisenhower, in his press release congratulating the Soviets scientists on Sputnik I, pointedly noted was separate from our military launch vehicle programs. President Eisenhower's private response to Sputnik was to increase the funding for an Air Force program first organized in 1953 that, after 12 failed attempts, resulted in the successful launch and recovery of the film carrying capsule of Discoverer XIV on 30 August 1959. The images were not as clear as the U-2 aircraft photos but covered Soviet areas never seen by spy planes. The purpose of the program was kept top secret until 1995 when President Clinton declassified it under its proper name of Corona. (See the article on the Military Use of Space in the Encyclopedia for more details.)

The Russians soon flew their own spy satellites, and in a kind of bilateral "don't ask, don't tell" spirit, the Russians finally, but tacitly, accepted Eisenhower's "Open Skies" policy. This policy actually followed the aircraft reconnaissance

focused Open Skies policy and is more accurately called the “satellite data non-discriminatory dissemination policy.” It supports the freedom of all satellites to image the surface of any and all countries by assuring access on an equitable basis by the imaged country to all images of its territory. (See discussion in the “Getting Data to Users” section.)

There were, of course, prophets. In 1946, well before Sputnik, the United States Army Air Corps requested that RAND Corporation consider how objects might be inserted into orbit (3). The study resulted in a report, Preliminary Design of an Experimental World-Circling Spaceship (4). The proposed midget moon, or “satellite,” would provide “... an observation aircraft [sic] which cannot be brought down by an enemy who has not mastered similar techniques.” The public sector had its own prophet. In 1951, 6 years before Sputnik 1, Arthur Clarke, a science fiction writer, proposed that a satellite could be inserted into orbit over the North and South Poles while Earth revolved beneath it, and that this satellite would permit humans to view the planet in its entirety (5). (Clarke is also credited with being the first to propose using satellites for communications.)

The civil satellite potential was not neglected. In April 1960, the National Aeronautics and Space Administration (NASA) and the Department of Defense (DOD) launched the Television and Infrared Observational Satellite (TIROS-I) into a polar orbit, inaugurating the first experimental weather satellite. This system generated the first television-like pictures of the entire globe in a systematic and repetitive manner. This ongoing series of TIROS satellites became operational in 1966 as the TIROS Operational Satellites (TOS) and in 1970, the National Oceanic and Atmospheric Administration (NOAA) renamed them the Polar Orbiting Environmental Satellites (POES). (For a full account, see the Weather Satellites in this *Encyclopedia*.)

In spite of the deep secrecy of the Discoverer Program, the military early recognized the need to involve the scientific community in developing techniques for using satellite data. In February 1962, at the Navy-sponsored *First Symposium on Remote Sensing of the Environment*, Dana Parker, one of the organizers, focused his inaugural address on the fundamentals of the electromagnetic spectrum, a subject that suggested that multispectral data could provide information that went beyond the “eyeball” analysis of photographic data on which all reconnaissance analysis was based. His talk heralded the multispectral interpretive approaches responsible for the selection of the Multi Spectral Scanner (MSS) flown on Landsat 1 and the current capability to analyze continental and global scale Earth cover changes (5).

The person who was probably the most influential in realizing and promoting the value of multispectral sensing was Robert N. Colwell, a professor of forestry at the University of California at Berkley. In the 1950s, Colwell and others showed that agricultural crops, trees, and even different soils possess a telltale electromagnetic signature that could be measured from any height and used to identify the object. He and his American colleagues were not alone in such thoughts. A Soviet scientist named E. L. Krinov had measured the spectral signatures of some 370 natural and man-made objects in the 1940s, and he, too, had aerial mapping in mind. In a prescient article in *American Scientist* in 1961, Colwell wrote, “Just as our musical appreciation is increased greatly when more than 1 or 2 octaves are exploited, so also is our appreciation of the physical

universe through multiband spectral reconnaissance, which already can exploit more than forty 'octaves'." He and his colleagues at Michigan and Purdue were instrumental in the selection of the MSS launched on Landsat 1 (2). Colwell's students were pivotal in making the rapid development of multispectral sensing possible.

Getting Landsat Started

The concept of a dedicated, unmanned land-observing satellite emerged in the mid-1960s. It was stimulated by the early interest of the ONR, and aided by NASA's studies and aircraft sensor experiments starting in 1963 in its Office of Manned Space Flight's Earth-Orbital Apollo Extension System under P. C. Badgley. These efforts bore direct fruit in Earth-sensing experiments on Skylab in the mid-1970s, but were more important to the Landsat decision because Badgley also funded the U.S. Geological Survey (USGS) and the U.S. Department of Agriculture (USDA) to define practical uses for space-derived images (6). By 1996, Badgley, then working in NASA's Office of Space Science and Applications, and Leonard Jaffe, in charge of its Application programs, were well aware of the desire of the USGS and USDA for the speedy development of an Earth resources satellite, but NASA upper management showed little inclination to address the desires of the potential user agencies. What attention they were giving to Earth application needs was centered on planning to use the Apollo spacecraft as a manned Earth observation system. (This should not be surprising. At this same time, the U.S. Air Force, in spite of the already very active unmanned observation program, was in the first stages of the Manned Orbiting Laboratory (MOL) program which would use the Gemini spacecraft to support a space station filled with both optical and radar sensors. MacNamara canceled it probably because of its rising costs, but also because there was no clear argument that the presence of a crew in orbit would add value to the reconnaissance data. He was also aware that the loss of a manned spy satellite due to enemy action would require a much more dangerous response on our part than the elimination of a simple satellite.)

While NASA's manned program developed and tested some of the relevant technology and the startling photographs from Gemini and later Apollo gave scientists insight into how valuable satellite imagery could be (7), the real credit for the Landsat satellite must go to the USGS and in particular to its then director, Dr. William T. Pecora. Disturbed by the continuing lack of progress by NASA, Pecora set up a press conference on 21 September 1966 at which his boss, Interior Secretary Stewart Udall, announced the Interior's plans for an Earth Resources Observation Satellite program (EROS) to be run in cooperation with NASA. James Webb, the NASA Administrator, made aware of the press conference on the previous day, arranged an immediate meeting with President Johnson, who reaffirmed that the development of space technology was NASA's responsibility. Secretary Udall was informed by NASA on the day after the press release that NASA had been assigned "lead agency" responsibility for civil experimental space applications. The USGS announcement had stressed its development of an operational system, but NASA managed to maintain that any

satellite system, even if it carried sensors considered ready for operational use, was experimental (6). This has been governmental and NASA policy to this day, though it has several times required some imaginative interpretation of “experimental.”

Although unsuccessful with its own “operational” satellite system, the Department of the Interior (DOI) continued to press NASA for a satellite. The overall goal of the proposed EROS program was to acquire remotely sensed data from satellites in the simplest possible way, deliver these data to the user in an uncomplicated form, and ensure their easy use (8). NASA did not believe that the problem was technically that simple and was especially negative about the schedule that the USGS was proposing. DOI seemed to have lost the battle, but the result was that NASA accelerated its planning for a small unmanned satellite.

However, it was not to be that easy. NASA was initially frustrated by the Bureau of the Budget (BOB) that rejected NASA’s FY 1968 budget. In the fall of 1967, NASA and the DOI budget request for FY 1969 funding for the project was again turned down. It took a direct appeal by NASA Administrator James Webb to President Johnson to have the request restored and a rescheduling of the elements in that budget by Congress to allow NASA to start on the initial contractor studies for the satellite. DOI’s budget request was still refused, which caused a delay in its plans to initiate a data processing and distribution facility. This delay would cause problems in getting Landsat 1 data to user communities for several years after the launch.

Though NASA finally was given its FY1970 budget request as well, the BOB was still fighting hard to stop the program; one internal BOB memo written in June 1968, while the FY 1970 request was being developed, suggested that the Landsat satellite proposal should be eliminated and replaced with an Earth resources aircraft program (6). Because the EROS satellite was being developed to prove that satellites images could be beneficial to a wide range of governmental and public activities, the BOB had required NASA to make a cost-benefit analysis of its potential value. The argument that aircraft could do the same jobs better and at less cost was an issue that plagued Landsat then and later as it presented its plans for the follow-on Landsats. Amron H. Katz, in an article in *Astronautics and Aeronautics*, made this argument most forcibly in June 1969, and he was still arguing the superiority of aircraft through 1976, long after the launch of Landsat 1 (6). At this time, Katz’s argument is in the process of being tested in the marketplace by the recently launched commercial satellites whose 0.6 to 1-meter resolutions can compete directly with many (but not all) of the imaging tasks performed by aircraft-based sensors.

Interior changed the “S” in its Earth Resources Observation Satellite (EROS) program from Satellite to System and placed it under the direction of the USGS. The EROS mission was to archive and distribute remotely-sensed data and to support remote sensing research and applications development within the DOI. To carry out the EROS responsibilities, the USGS built the EROS Data Center in Sioux Falls, South Dakota, in 1972. This location was chosen after a competition with a location in Mississippi. Both were being considered because their central locations would enable a receiving antenna to record satellite passes over both the East and West Coasts. The receiving antenna was finally installed

at the EROS Data Center (EDC) in 1998, the delay being due to the commercialization of the program since the contractor chose Norman Oklahoma as the location for its receiving station.

The Multispectral Scanner Years, 1972 to 1984

Table 1 provides the complete history of the Landsat system, its sensor complements, and operational periods.

Landsat 1 carried two sensors, the return beam vidicon (RBV) and the MSS. The RBV was a television camera designed for cartographic applications and was the sensor of choice of the USGS and originally NASA. The MSS was designed to identify natural features using spectral analysis and came to NASA in an unsolicited bid from a group at Hughes Aircraft headed by Virginia Norwood. The Hughes scanner triggered quite a debate. “Mapmakers like myself were very suspicious of the multispectral scanner, which we could not believe would have geometric integrity,” admitted the USGS’s Alden Calvocoresses. “We were wrong on that one.” “People were so emotionally opposed to a mechanical device,” Norwood recalls. The argument which went on for more than a year, was resolved by flying both. However, the six-band sensor originally proposed by Hughes had to be reduced to a four-band system because of weight restrictions. In retrospect, the decision was to prove a crucial one. Soon after the launch, the TV cameras became afflicted with an unexplained electrical problem and had to be shut down within the month. The scanner, designed to work for 1 year, was still functioning, moving mirror and all, when the satellite was turned off 6 years later. Figure 8’s schematic of the MSS shows how the mirror scans across the track while the satellite motion provides the other dimension. Figure 9 shows the whole satellite with the locations of the MSS and RBV noted.

The launch of Landsat 1 in the summer of 1972 settled the issue. The quality of the images, including their geometric fidelity provided by the multispectral scanner (MSS), was a surprise to many in the community and was quickly followed by the recognition that its data scale, resolution, and four color bands provided new, unique, and very useful ways to see and understand our geography.

Table 1. History of the Landsat Satellites

Satellite	Launched	Status as of July 2002	Sensors
Landsat 1	7/23/72	Ended ops 1/06/78	MSS & RBV
Landsat 2	1/22/75	Ended ops 2/25/82	MSS & RBV
Landsat 3	3/05/78	Ended ops 3/31/83	MSS & RBV ^a
Landsat 4	7/16/82	Standby	MSS & TM
Landsat 5	3/01/84	Operational	MSS & TM
Landsat 6	10/05/93	Launch failure	ETM
Landsat 7	5/15/99	Operational	ETM +

^aLandsat 3 was kept “experimental” by adding a thermal band to the MSS and by replacing the three 80-meter RBVs with two 40-meter panchromatic cameras. The thermal band failed, and the added resolution of the RBVs did not provide sufficient improvement to increase their low use relative to the MSS.

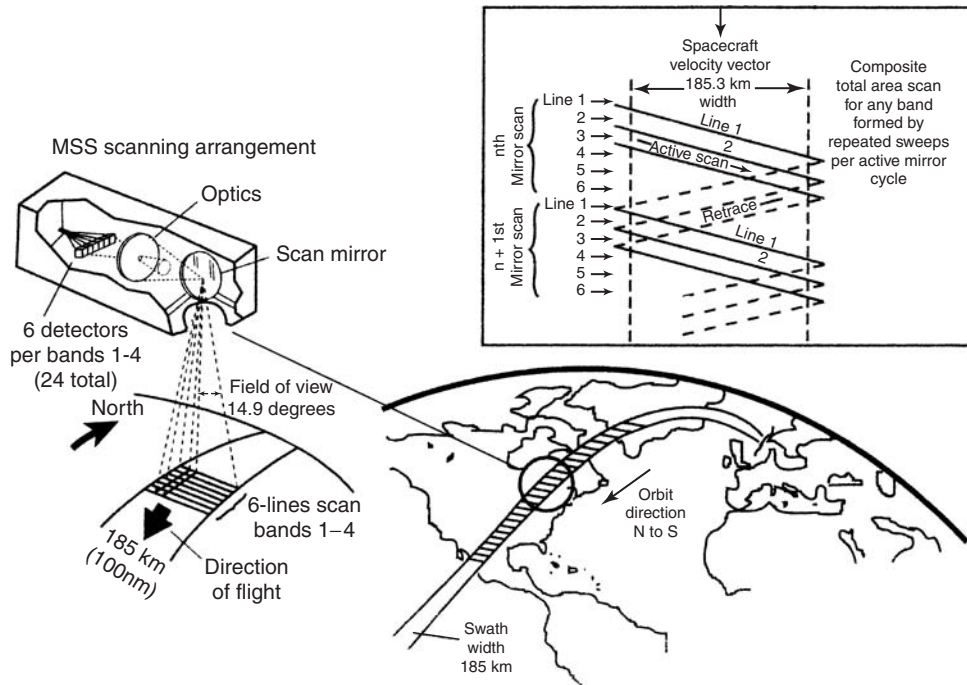


Figure 8. Scanning geometry of the multispectral scanner (from Landsat 4 Data Users Handbook, 1984, USGS).

The satellite data were received on Earth in digital form. However, nearly all of the initial MSS products were provided as color and black and white prints and negatives because the user community had little experience with or training in analyzing data in its digital form. In addition, their computer facilities were only marginally up to the job of processing the amount of data involved, and the interpretive algorithms were still in an early developmental stage. There was a cost difference that certainly had some effect in the later years; film costs varied between 8 and 15% of the cost of digital tapes. Film dominated the product sales in dollars until 1984, but even then 35,000 film scenes were delivered compared to 5,000 digital scenes. Digital sales increased rapidly after that until, in 1993, EOSAT eliminated its photographic product line because it was not economically viable (9). Finally, 21 years after Landsat 1, the digital applications made possible by the MSS, had worked their way into the marketplace. Today, all of the 24,000 annual sales are in the digital format.

The quality of the photographic images was more than sufficient to inspire a rush to explore their uses for many applications. For the first time, mapmakers had images that covered 100 miles on a side with near orthophotographic geometry. (The edge of the scene was imaged at only $7\frac{1}{2}^\circ$ from the vertical. Figure 10 shows just how impressive this capability is. It is one Landsat scene covering $185\text{ km} \times 170\text{ km}$, virtually 100 miles on the side, showing the Chesapeake and surrounding land, including Baltimore and Washington, D.C. Aircraft montages for that same area would involve hundreds of

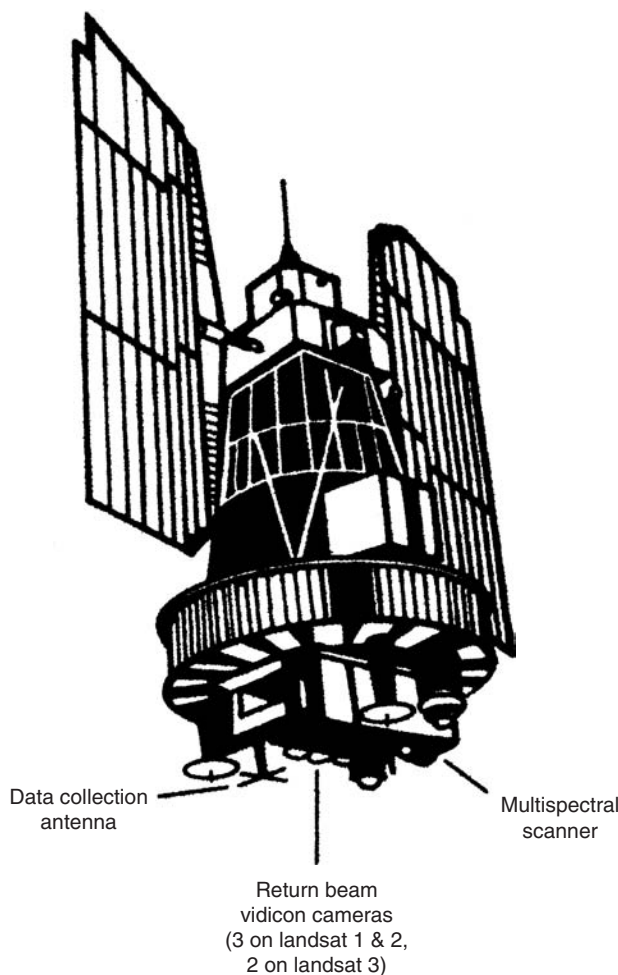


Figure 9. Platform configuration, Landsats 1, 2, 3.

individual images, each of which would have had to be corrected to match their 30 or 40° edge angles with their neighbors.) Geologists were able to recognize innumerable new fault lines revealed by the consistency of the lighting and shadow angle across the whole scene. There was even a minor coffee-table book rush to present the public with pictures of both familiar and far away places made strange and beautiful by the many colors generated by manipulating the four color bands.

The first sizable use by industry was by the major mineral and oil exploration companies. They were targeted early in the program by JSC's Earth Resources program office and, after some initial skepticism, formed Landsat analysis groups that used Landsat images to update maps and to identify surface features that warranted further exploration. It was equally important for them to be able to obtain images anyplace on the globe without their rivals knowing about it. (Early in the program, EDC established the policy of keeping the data



Figure 10. 100 mile square Landsat image of Chesapeake Bay and environs. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

buyers' identities private.) The industry put together an advocacy group, the GEOSAT Committee, that served as its representative in the many Landsat debates that were to be a feature of the Landsat program over the years. Robert Porter, who resigned his position as Director of the NASA Earth Resources Program in 1970 to found his own company, Earth Satellite Corporation, to apply Landsat analysis to commercial needs, estimated that at least \$1 billion worth of oil had been found using Landsat data (6).

The boost this application gave to those in NASA who defended the Landsat program would be short-lived; the exploration companies generally needed only one look. Landsat covers the whole globe 24 times a year, every year, and its "killer application" must be one that needs that kind of repetitive coverage. The management of things that grow, trees, grass, and food, seemed to fit that bill and were, of course, the intended target of the multispectral analysis for which the MSS was designed.

Transferring Landsat Technology

NASA's Office of Space Applications (OSA) was split from the Office of Space Science and Applications (OSSA) in 1971 and retained the Landsat program that had been transferred to OSSA from the Office of Manned Space Flight in 1966. OSA also developed the first communication satellite, the first weather sensors and satellites, both polar and geosynchronous and the first and still the only U.S. civil radar satellite, Seasat (See Radar section). However, it recognized from the start that its function went beyond the development of satellites and included the development of the required ground systems and data analysis technologies that

would enable users to apply the satellite data to their problems. OSA's success in these basically technical problems can be rated excellent for the satellites and sensors and after a very slow start, good for the ground systems and analysis algorithms. But it soon was forced into the social and political engineering role of technology transfer. OSA's results in this function are still being debated. The best study of this activity and its results for the period up to 1990 is Pamela Mack's *'Viewing the Earth: The Social Construction of the Landsat Satellite System'* (6), which has been the source of many of the facts in this article. As will be noted later in this article, the social and political issues concerning Landsat and its successors became even more complex after 1990.

NASA's initial approach to the technology transfer issue was to fund studies in universities and user agencies. The university studies were most often concerned with developing and testing analytical techniques to exploit the manipulation of the signals from the four bands to identify crops and to measure their yield and health. The user agencies were concerned about defining operational and land management functions that could be done better, and more economically using Landsat images, or new but needed applications that could not be done at all without Landsat. NASA requested proposals even before Landsat 1's first launch. Expecting 40 or so principal investigators, it was inundated by more than 500 proposals. Wanting to encourage the broadest use possible, it set up peer review panels in several disciplines (chiefly geology, hydrology, and geography). The panels chose more than 300 investigators (6).

The large number of investigators made NASA carefully consider what the data distribution policy should be. Because there was the possibility of using an image to find value on someone else's property, NASA decided to release all data to the public immediately to avoid charges of unfair discrimination. Data distribution policy is a crucial issue still facing the civil program, involving as it does the complex relationship between the commercial interests that require exclusive distribution to earn their profits and the government's desire to provide data to scientists and government agencies at prices commensurate with "public good" support.

The results of these and following studies were presented in a series of symposia hosted by GSFC starting in March 1973. The presenters were asked to report on "user identified significant results" in agriculture/forestry, environment, geology, land use/land cover, and water. Each of the *Proceedings* approached 2000 pages. The studies produced valuable scientific results from the time of the first launch, but they also revealed that some of the capabilities promised by the magic of multispectral analysis fell a bit short when faced by the heterogeneity of the real world and the radiometric and resolution limitations of the MSS. Though many crops could be identified, the accuracy of the identification varied widely by crop type, growth status, atmospheric conditions, and field size. There was much to learn, but the reports demonstrated the value of multispectral analysis for a wide variety of surface analyses. These analyses were also able to identify the limitations on analysis imposed by the limited resolution and number of spectral bands of the MSS, thus giving support to the need to develop a better scanner. The development of the new scanner, the Thematic Mapper, is described below.

The Large Area Crop Identification Experiment (LACIE)

The NASA technology transfer effort did not rely only on the numerous individual studies that NASA continued to fund throughout this first period. In the 1970 and 1971 growing seasons, the Johnson Space Center (JSC), the USDA, and Purdue University used aircraft equipped with a variety of sensors, including a multispectral scanner that simulated the MSS, to fly over 210 test areas in seven Corn Belt states. The program detected the corn blight early enough to check it by pesticides and, as a result, led to the operational use of low flying aircraft to solve the farmers' problem (6). It did not provide Landsat with a direct application, but it did provide valuable experience in dealing with the problems of crop identification and health detection in the real world, which were important in developing NASA's Large Area Crop Inventory Experiment.

In July 1973, Congress asked NASA why it had failed to do more to predict the failed Russian wheat crop of 1972. Congress was concerned because American farmers lost potential profit by being unaware of the condition of the Russian crop. NASA saw an opportunity to reply with a plan when Robert MacDonald, who had led the corn blight team at Purdue and was then the Landsat Chief Scientist at JSC, suggested the large-scale agricultural experiment that later became LACIE. LACIE was certainly large scale; it monitored the entire Russian wheat crop and the U.S. Great Plains for three years, involved a contractor force of more than 200 Lockheed analysts, and cost between \$10 and \$15 million a year. In 1997, its third year, it met its goal of 90% accuracy, successfully predicting the poor Russian wheat harvests in that year. But as a technology transfer mechanism, it failed. USDA, in response to the direction of the OMB (see below) formed the ArgriStar program in close cooperation with NASA, but chose to focus it on improving and testing the techniques of crop identification rather than continuing the country production estimates of the LACIE program. Over the many years since, it has developed new and different uses of Landsat data and has been the largest civil agency Landsat data purchaser.

The End of NASA Application Activities

LACIE's last year was 1997, as it was for most of OSA's Landsat technology transfer programs. The Office of Management and Budget (OMB, the new name for the BOB) informed NASA that it would no longer support NASA activities that benefited other agencies on the theory that if they were really beneficial to those agencies' functions, the agencies should budget for them. NASA was being told to return to its experimental and scientific mission. This was part of the commercialization of the Landsat drive discussed in the next section. By the end of 1978, OSA had virtually ceased all application efforts, replaced its application managers with science managers, and embarked on studies to define the scientific requirements for Earth observation satellites. This effort resulted in the creation of the Office of Earth Sciences and the development of the Earth Observation Systems (EOS) program designed specifically to support the interagency Global Change program. Landsat was eventually to be considered part of this program, but it received little

management attention, which resulted in little science or application research funding during Landsat's commercial phase. NASA did, however, continue its development of the Thematic Mapper throughout this period, as befitted its technology development role (11).

The Thematic Mapper

As noted before, during the 1972 to 1978 period, NASA developed a new and improved multispectral sensor, named the Thematic Mapper (the TM) to emphasize that it was designed to measure and map the basic features or themes of the earth's surface automatically, that is, forests, plains, rivers, lakes, cities, farms, etc. Originally the sensor designers at GSFC reacted to the success of the MSS by proposing two alternate improvements, a 10-meter high-resolution pointable imager (HRPI) and a 30-meter resolution TM. HRPI was to be a small sensor using a new technology, multilinear array sensors, the scanner technology adopted by the French for their SPOT system in 1978 and for virtually all (except Landsats 6 and 7) land imaging sensors since then. HRPI could be programmed from the ground to point to targets along or to the sides of the flight path. This pointability provided 2 to 3 day repeat looks at a given target. It also made it possible to generate stereo images. However, high resolution came at the price of a small field of view, 10 kilometers. The decision between the two was a fairly easy one for NASA because the scientists liked the large area coverage of the MSS and DOD made it plain that 10 meters exceeded their threshold for civil system resolution. It is interesting to speculate on the path remote sensing would have taken if HRPI had been chosen. HRPI had the pointability and higher resolution that have been the capabilities that finally emboldened the private sector to fund commercial systems in the mid-1990s.

The TM was selected because its 185-kilometer wide, 16-day global repeat coverage replicated that of the MSS, a very important requirement of the science community and user agencies. The user agencies and the scientific community were called upon to determine the exact spectral bands for the TM. The consultation process culminated in a 3-day conference at Purdue under the direction of Dr. David Landgrebe, the director of their remote sensing group. Three teams, each composed of sensor designers, data analysis experts, and data users, debated the merits of all possible band combinations. There was surprising consensus among the three teams, and the selection of six bands was essentially unanimous (12). Much later in the sensor development process, the geological community led by Alex Goetz, feeling they had been slighted in the band selection process because of NASA's focus on agricultural users, requested that a band he had shown was sensitive to minerals, be added. Fortunately, the addition was found technically feasible, and a seventh band was added. The increased capabilities of the TM came with significant weight and size increases over the MSS. The TM weighed 258 kg versus the MSS' 64 kg and it measured $1.1 \times 0.7 \times 2.5$ meters compared to the MSS' $0.5 \times 0.6 \times 1.3$ meters. Figure 11 presents a picture and cutaway of the TM.

The user community also was able to convince the OMB that any satellite carrying an experimental sensor should also carry the MSS to allow testing the

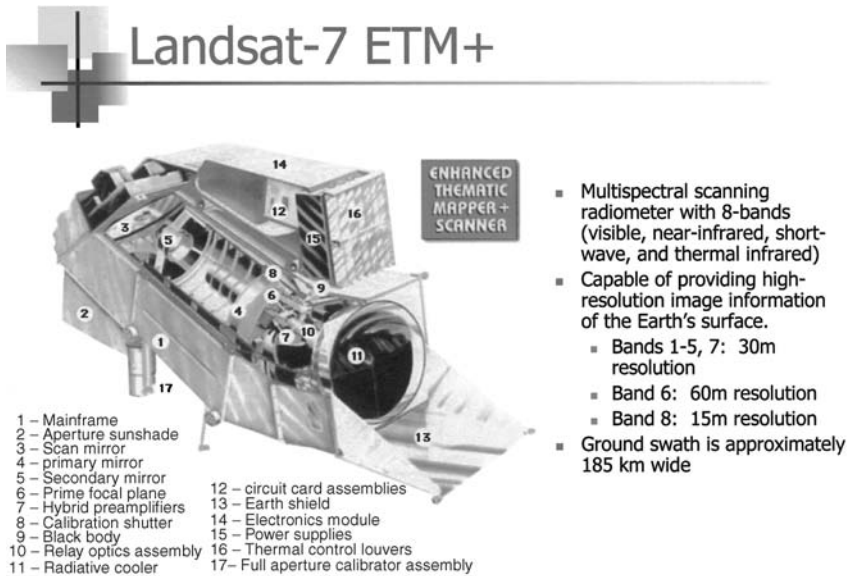


Figure 11. Landsat-7 ETM+ (note dimensions are $1.8 \times 7 \times 2$ meters). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

new sensor relative to the established and understood characteristics of the MSS. While this was a scientifically reasonable position, the user agencies were also worried that they would have budget problems upgrading their computer systems to meet the greatly increased data volume generated by the TM. Thus, Landsats 4 and 5 came into being. Figure 12 presents a drawing of Landsats 4 and 5.

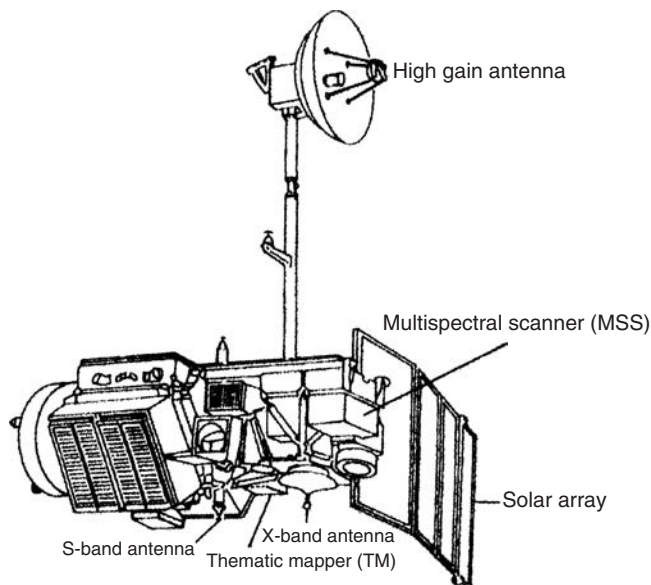


Figure 12. Landsat 4 and 5 observatory configuration.

Commercialization 1984–1992

When the TM on Landsat 4 was launched in 1982, NASA could declare Landsat a technical success. Landsat was also an international success as the look-alike programs of France, India, Russia, and Japan were soon to demonstrate. However, as is the fabled fate of prophets in their own land, NASA failed to convince any other agency or any commercial entity to develop its own system. The user agencies, which were vociferous in telling NASA what they needed to make the program better and, above all, their need for it to be operational, were very cautious in their testimony to Congress on the value of Landsat, lest they be asked to use part of their budget to build the next system.

The basic issues had been there from the start: should there be a civil operational Landsat and which agency should fund and manage it? The large benefit–cost ratios derived from the BOB-requested economic cost studies and used by NASA to justify the continuation of the program, inadvertently added a third issue: should the system be commercialized? In June 1978, the Office of Science and Technology Policy published the results of a study of policy options for civilian remote-sensing satellites. The study concluded that the President should make a commitment to an operational Earth resources satellite system and suggested that NASA have responsibility for that system (6). User agencies and the states disagreed that NASA be the manager. In November 1979, President Carter issued Presidential Directive 54, which gave NOAA temporary responsibility for operational Earth satellites and called for the National Oceanic and Atmospheric Administration (NOAA) to develop a plan for turning over Earth resources satellites to private industry. The decision to have an operational remote sensing system and to have it managed by NOAA along with its management of weather satellites was roundly applauded. However, the suggestion that it should later be turned over to private industry sparked a strong controversy that has continued to this day. A 1980 report by a task force of state representatives explained that “the establishment and operation of a land remote sensing satellite system is viewed as a public service in the same context as census, cartographic, geological and meteorological data which are supplied by the Federal government” (Recommendations of the National Governor’s Association April 30th 1980). As required by the directive, NASA turned over the operations of Landsat 4 in 1982 after assuring that it was functioning as designed.

Carter’s steps toward gradual privatization were quickly trumped by Reagan whose administration moved rapidly to transfer the Landsat program to the private sector. Reagan attempted to include the weather satellites in this process but strong objections from Congress and a scandal concerning questionable lobbying and contact between COMSAT (which lobbied strongly because the privatization mechanism was based on its successful commercialization of communication satellites) and the Department of Commerce resulted in PL 98-52 which prohibited the transfer of weather satellites to the private sector. Reagan signed the final law, Public Law 98-365, the Land Remote-Sensing Commercialization Act of 1984.

In January 1984, NOAA solicited bids to operate the existing Landsats, retrieve and develop a market for their images and, aided by government

subsidies, to build and operate future systems. Both of the two final bidders, EOSAT a partnership between Hughes, the MSS contractor, and RCA the satellite contractor, and Kodak, required government subsidies of about \$500 million spread out over a 6 to 10-year period. This was a reasonable amount based on several NOAA-funded studies that agreed it would take at least 10 years before the sales of Landsat images could support the development of a satellite. OMB did not want to supply federal funds. Reagan's chief of staff split the difference, and the contract was limited to \$250 million of government funding. Kodak withdrew, and EOSAT was selected. However, the cost cap kept contract negotiations going until both parties agreed to a final amount of \$350 million, and the contract was signed in September 1985.

With that signing, the Landsat community breathed a sigh of relief because it seemed that Landsat was finally on a course toward the long desired goal of operational status. This was short-lived, however, as the impact of the EOSAT 200% price increases and their restrictive licensing requirements became felt. This was especially critical for the academic community and the state users who were used to buying the data at almost the cost of reproduction. The result was that during the commercial period, studies to develop and test new uses of multispectral analysis using Landsat scenes almost came to halt. Most of the major advances in spectral analysis of satellite imagery in this period were made using the free 1-kilometer resolution data from the advanced very high resolution radiometer (AVHRR) on the NOAA weather satellites.

The Reagan administration continued its campaign to reduce spending on Landsat even after the contract was signed; this made it virtually impossible for EOSAT to continue its marketing efforts and to plan for the follow-on satellite. In 1986, the French government launched SPOT1, designed specifically as a commercial data gathering system. EOSAT was faced with selling 30-meter data against SPOT's 10-meter panchromatic and 20-meter color data. In 1991, Landsat's sales were \$32 million and SPOT sales topped \$40 million. Note that neither number approached a level that would support a fully commercially financed system.

The first positive boost to the program occurred early in 1989 after NOAA directed EOSAT to turn off Landsats 4 and 5 due to lack of program funds. The newly elected Bush Administration was deluged with hundreds of letters from academic researchers, foreign governments, private companies, and concerned congresspersons demanding the necessary funding. Vice President Quayle, head of the newly formed National Space Council (NRC), canvassed the major user agencies and put together an ad hoc funding plan that kept the satellites in orbit. Unfortunately, there was still division in the administration, and the ad hoc process had to be invoked for the next 2 years. Landsat did start to gain backing in the government due to three factors: Bush's Global Environmental Change initiative, the utility of Landsat and SPOT data during the Gulf War; and the technological impact of computer capabilities, especially Geographic Information Systems (GIS) and the satellite-based Global Positioning System (GPS). The first had impact because Landsat was the only consistent set of global data extending back to 1972, and the second because it caused the powerful military and intelligence communities to state openly the importance of Landsat for their missions. (By late 1980, the DOD was the single largest user of Landsat data in the

world.) The third factor finally allowed science and application users to take full advantage of the digital data on affordable computers. The forces for change were set in motion.

Landsat Returns to the Government in 1992

Landsat commercialization was a failure. That is not just an opinion; Congress so declared it in the findings of the Land Remote Sensing Policy Act of 1992. Passed in October of that year, it was the result of a congressional process that was started by Representative George Brown and Senator Larry Pressler. Brown, Chair of the House Science and Technology Committee, was a dedicated environmentalist, and from that viewpoint, he declared that Landsat should be returned to the government to be treated as the weather satellites, as a public good (13). He introduced a bill in the house for the purposes of “Amending the Land Remote Sensing Commercialization Act of 1984...” and Pressler offered a similar bill in the Senate.

The Land Remote Sensing Policy Act of 1992 was a farsighted and fruitful act. It created the commercial licensing process that culminated in the current leadership of the United State in the high-resolution commercial market. It recognized the importance to this country’s scientific, resource management, security and economic sectors of continuing the then 20-year recording of the global land surface by the midresolution Landsat series and established “continuity” as the criterion for Landsat follow-on systems. That somewhat loosely defined criterion resulted in Landsat 7 and is the basis of the current plans for its successor, the Landsat Data Continuity Mission (LDCM). The law also mandated that the data be provided to everyone for the “cost of fulfilling user requests,” responding to Chairman Brown’s desire to make Landsat data a public good. It gave the DOI the responsibility for creating and maintaining an archive of land image data, a responsibility that the DOI assigned to the USGS. Subsequently, the USGS entered into an interagency agreement with NASA to operate Landsat 7 and its data collection, archiving, and distribution functions. The law recognized those desiring commercialization by giving the DOC, appropriately consulting with the DOD, CIA, and the NRC, the task of licensing private companies to fly Earth-sensing satellites.

The administration was not idle while Congress debated. In March 1992, NASA and DOD released the management plan that Bush had requested. The plan called for the development of Landsat 7 to replace the EOSAT-built Landsat 6 at the end of its expected life and to continue the “continuity” of the Landsat data series. The plan assigned the responsibility of the space segment for Landsat 7 to the DOD and the ground segment to NASA and provided a budget of \$470 million for DOD and \$410 million for NASA. After so many years of cross-purposes, it seemed that the Bush administration and Congress agreed at last on the role of civil remote sensing satellites.

The real breakthrough for the commercial satellite companies occurred as the result of a joint meeting of the House committees on Science and Intelligence in 1994. One company, Earth Watch (now Digital Globe), quickly obtained a license for a 3-meter resolution satellite. (Unfortunately, it failed soon after

launch in 1997.) However, at the time of hearing, Lockheed had been waiting 6 months for an answer to its request for a 1-meter system. Congress wanted to know why. The testimony of all of the agencies was that the availability of 1-meter images to the world might pose a serious threat to our security and required more study. Congress suggested that they reevaluate their concerns in the light of the 2-meter satellite scene of the Mall that had been purchased on the commercial market from Russia. Two weeks later, President Clinton issued Presidential Decision Directive (PDD) 23 that authorized commercial satellite high-resolution imagery and set specific regulatory guidelines. It left the definition of high resolution to the decision of the agencies involved, but its intent was clear to all, and Lockheed had its 1-meter license. At the present time, two companies have licenses for systems of 0.6-meter resolution and one is already operational.

Though the DOD overplayed its hand a bit, it did have legitimate concerns about the availability of such data in time of war. The license contains a “shutter control” clause for just that situation. However, shutter control’s legality is being questioned on the basis that it is prior constraint and not allowed by the Constitution. During the Afghanistan engagement, the National Imaging and Mapping Agency (NIMA) finessed the prior constraint issue by contracting with SpaceImage for exclusive use of all the data taken over Afghanistan during the critical months of the operation. Most of the data were made available to all 2 months later.

Landsat Returns to NASA

Hardly a year passed before the Landsat program was evaluated for a third time, principally because NASA felt that its budget would not cover building the extra capabilities in the ground data system that would be required by DOD’s addition of a high-resolution (5 meters) multispectral stereo imager (HRMSI). The National Science and Technology Council (NSTC) meeting in response to the launch failure of Landsat 6 and aware of DOD’s reluctance to continue a program without a HRMSI sensor, recommended developing Landsat 7 with only an improved TM instrument and establishing a new management structure, so that DOD could withdraw from the program. This resulted in Presidential Decision Directive/NSTC-3, dated 5 May 1994, reconfirming the Administration’s support for the program but giving NASA, NOAA, and the USGS joint management responsibility (White House, 1994). The Landsat Program Team (the LPM) was created by the three agencies. It proceeded to negotiate with EOSAT for new Landsat 4 and 5 product prices for the U.S. government and its affiliated users and began the process of defining and developing Landsat 7.

A project office was established at GSFC. It defined a new sensor, the ETM+ (extended TM+), that replicated the TM bands but also included several improvements made possible by advances in technology since the earlier Landsats. The improvements included adding a 15-meter panchromatic channel, sharpening the thermal band to 60 meters, and increasing sensor sensitivity and calibration accuracy. It also defined a data acquisition goal. The system was to be

able to acquire the entire landmass of the globe on the average of four times a year. Thus, for the first time in Landsat history, the satellite was to be programmed to do what its supporters had long envisioned: acquire a scientifically useful annual archive of the seasonal and annual changes that take place on all of Earth's land surface.

Landsat 7 was launched on 15 April 1999. Its data have met all expectations. Landsat 7 is the first and still the only midresolution system programmed to acquire and archive all of Earth's land surfaces once per season. The data acquisitions are being collected under the guidance of a program developed by Dr. Sam Goward's team at the University of Maryland. The program maximizes the probability of observing critical seasonal changes and reduces the number of cloud-filled scenes acquired from the system's capability of 250 scenes per day. This is not a trivial task because clouds cover 50% of the globe, on average, and some of the most biologically active areas are covered 80% or more of the time. (For the past 2 years, scientists have been able to gather data at twice the frequency of the 16-day Landsat repeat orbit because Landsat 5 is still functional in its eighteenth year of operations.)

Equally important, especially to the global environmental change science community, as required by the 1992 law, all of the data are archived at EDC and are sold to everyone at the cost of fulfilling user requests, currently about \$600 per scene for the standard product. A survey of EDC image sales for the first 10 months of Landsat 7 operations showed that sales averaged 2400 a month. An analysis of the 1000 separate purchasers revealed that when account is taken of the many commercial and academic buyers who purchase data for use on government programs, the government is the ultimate purchaser of more than 80% of the data sold. This supports the observation in the 1992 Law that Landsat data are effectively a public good.

The Landsat Data Continuity Mission (LDCM)

Goddard created a Landsat Data Continuity Mission (LDCM) team with the USGS in 2000 (NOAA had left the Landsat Program Team earlier because it did not get its Landsat budget) to formulate plans for a follow-on mission to Landsat 7. In response to Congressional mandates that NASA not develop any satellite until it could demonstrate that it could not buy similar data commercially, GSFC first developed a rigorous set of data specifications that would meet the requirement for data continuity: 30-meter multispectral resolution, 15-meter panchromatic, 185-kilometer swath, data quality equal to or better than Landsat 7, and delivery of an average of 250 scenes a day for 5 years. After several iterations with the science and application community, the resulting data specifications were made the basis of an RFP to industry for a 6-month study to provide several trade-off studies to NASA, including the cost of keeping the thermal data requirement. The contractors were also requested to provide the design of a total satellite, sensor, and ground system to the level of detail required for a traditional Program Design Review. Two contractors were chosen. Both bids were for systems designed to supply a commercial market that required much higher capability than NASA was requesting, but whose data could be degraded to meet

the government's requirement exactly. It is just possible that a formula has been found that will meet the requirements of both sides of the commercialization–public good debate. NASA will get a product that meets its scientists' goals without having to remain in the satellite operations business, Landsat type data will be provided to all at the cost of reproduction, and industry will get the help it needs to develop new products that will broaden and add to its nascent high-resolution business.

Getting Data to Users

The Landsat history described so far left out a major element in the Landsat program; the ground system hardware and software that received, processed, stored, and distributed the data to users. The history of the development and operation of the ground-based systems is as full of trials, tribulations, decisions, and revisions as there are in the space segment's story. The facts in the following summary have been taken (sometimes literally) from an excellent article by W. C. Drager et al. (9).

To Pecora (8), the success of Landsat depended on making the data available to users in a timely and efficient manner. NASA recognized that it was its responsibility to develop the hardware and software to process the data from its new instruments. The technical problems turned out to be more difficult than expected, and data processing systems were slow in reaching the ability to handle the amount of data that was being downloaded from the spacecraft and ordered by users. Also, NASA did not think that its job was to be operationally responsible for the processing, storage, and distribution of the data products. Thus prior to the launch of Landsat 1, NASA signed an agreement with the USGS to process, archive, and disseminate Landsat data at the USGS EROS Data Center (EDC). That agreement, together with similar agreements with NOAA during its Landsat tenure, was codified by The Land Policy Act of 1992 that directed DOI/USGS to maintain a national archive of land remote sensing data and, as noted before, to make it available to all users at the cost of fulfilling user requests. That remains the situation currently.

As the United States established the Landsat program, it made extraordinary efforts in the United Nations and elsewhere to gain worldwide acceptance of nondiscriminatory dissemination of remotely-sensed data from space. The United Nations adapted this position in 1986 as "The Principles Relating to Remote Sensing of the Earth from Space" (10). This concept (often called the "open skies" policy, although that name is more correctly applied to an Eisenhower initiative to exchange the rights to do aircraft remote sensing with Russia and other nations) was responsible for many early activities aimed at getting Landsat data into the hands of data-poor third-world countries. Soon after the launch of Landsat 1, U.S. development assistant agencies, including USAID and the International Development Bank (IADB), began offering sizable Landsat data grants and funding for personnel to third-world countries. In 1974, for example, USAID awarded \$260,000 worth of Landsat data and training grants to 10 developing countries in South America and Africa. By 1981, that program had grown to approximately \$40 million and included projects in

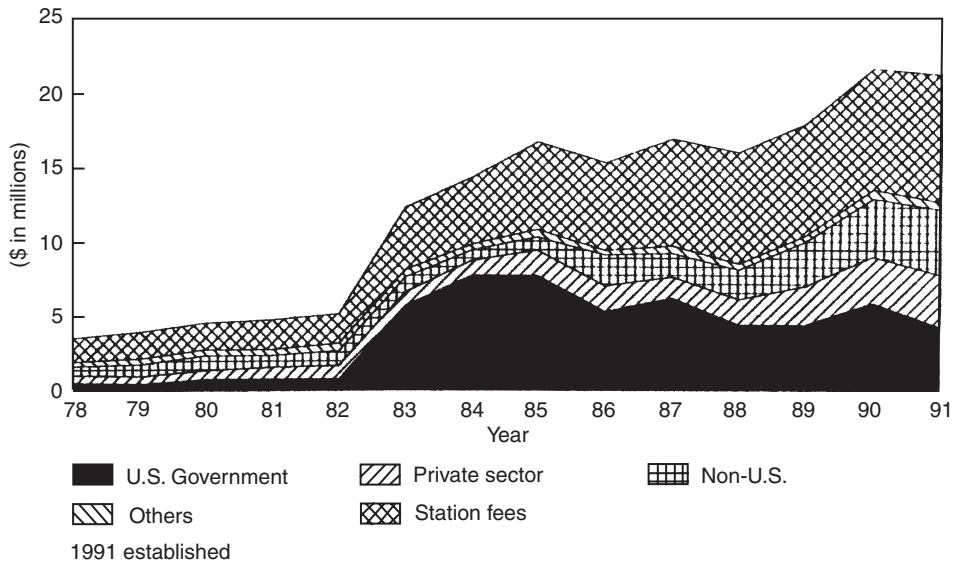


Figure 13. Landsat revenue history.

35 countries that ranged from forest surveys in Costa Rica to geological mapping in Monaco (11).

NASA and the USGS initiated another major "open skies" initiative, the establishment of Landsat data receiving stations in foreign countries. NASA agreed to downlink Landsat data directly to such stations every time the satellite passed within its range. In addition to disseminating the data to the widest possible user audience, the establishment of foreign ground stations had two other very practical reasons: they served as a very valuable backup for the failure of the onboard tape recorders that were the weakest link in the data system, and they provided a significant source of revenue to NOAA that offset its operational expenses. The early annual fee of \$250,000 rose to \$600,000 by 1998 (11). Figure 13 presents Landsat's revenue history that shows the large part that station fees played. Each station has the right to sell all of the data it receives and the responsibility to archive that data for possible use by scientists; this responsibility has not been followed to the extent that the Landsat scientists would like. However, the foreign stations have been very active in distributing Landsat data. From 1979 through 1995, more than \$234 million worth of data has been distributed worldwide; at least \$103 million of that was through foreign ground stations (9).

Starting with Canada in 1972, 17 countries have built Landsat receiving stations over the years. The names and locations of 16 stations are presented in Fig. 14. (Argentina, Spain & Taiwan created stations after the period shown on the chart.) In the midlatitudes, each station covers an area the size of the United States, and at the higher latitudes, a great deal more. The result is that the ground stations cover virtually all of the global land cover except a large portion of northern Asia, principally Russia and China, and central Africa. These stations, all of which are government run or supported, have become their country's focal

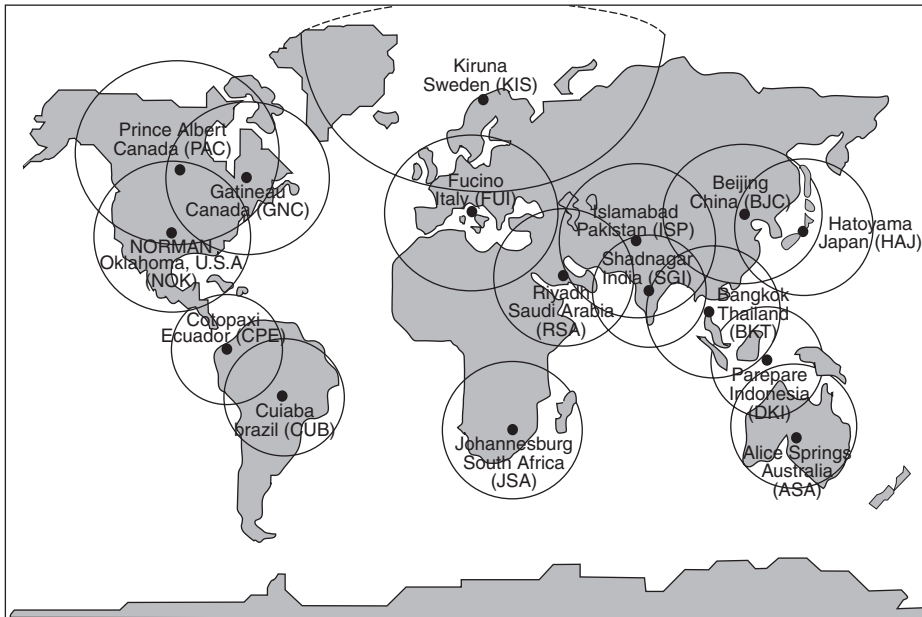


Figure 14. An example of Landsat receiving stations and areas of coverage.

point for remote sensing. In recent years, many have established agreements with SPOT, Radarsat, and commercial high-resolution satellite operators.

There is no question that the Landsat program's early and continued aggressiveness in promoting and supporting foreign ground stations has been the major element in developing the worldwide application of satellite data and of creating the many national infrastructures that are critical in exploiting the satellites' unique global capabilities for both global science and the creation of a global market for commercial systems. This activity also became important to the Landsat program itself; the support of the international ground stations aided in the many debates on whether Landsat should be continued.

The International Period

The "United States only" period ended rather abruptly in late 1985 as the foreign activities generated by the early Landsats culminated in the launch of multispectral satellites by Russia, France, Japan, and India, all within the following 2 years. Though it took a little longer, Seasat's radar images also impacted the international activities starting with the launch of Russia's COSMOS-1870 in 1987, followed a year later by the Japanese and ESA launches.

This "international" period of very aggressive foreign development and launch of civil land-imaging satellites extended through 1996. During this second period, France, India, Japan, and Russia launched 14 optical systems. The United States' only launch, Landsat 6, did not make orbit. It was only the incredible extension of the orbital operations of Landsats 4 and 5 beyond their

Table 2. Resolution, Bands and Swath Width of Some Current Satellites

Satellite	Orbit	Sensor types	Resolution in meters										SW Km	
				Thematic mapper bands										
			Pan	VNIR				SWIR	MWIR	TIR				
				1	2	3	4	5	7		6			

Landsat like, frequent global coverage

IRS-1 C	'95	M&P	5	23	23	23	70						70, 142
IRS-1 D	'97	M&P	5	23	23	23	70						70, 142
Spot 4	'98	M&P	10	20	20	20	20						120 ^a
Spot 5	'02	M&P	2.5, 5	10	10	10	20						120 ^a
CBERS-1	'99	M&P	20, 80	20	20	20	80	80			160		120
Landsat 7	'99	M&P	15	30	30	30	30	30			60		185

High resolution, small area coverage

Ikonos 2	'99	M&P	1	4	4	4	4						12
QuickBird 2	'01	M&P	0.6	2.5	2.5	2.5	2.5						16
Eros-A1	'00	P	1.8										13
IRS TES	'01	P	1										?

Multi & hyperspectral experimental

Terra	'99	M		15	15	15	30	5 @ 30			5 @ 90		60
EO-1	'00	H&M		233 bands @ 30									15
MTI	'00	M		4 bands @ 5, 3 @ 20				3 @ 20	2 @ 20	3 @ 20			13
PROBA	'01	H&P	8	19 @ 8 or 63 @ 36									18

Radar

Radarsat 1	'95	R	8.5		50
ERS-2	'95	R	30		100
ENVISAT	'02	R	30		100

Companion low res wide area sensors

Sat	Inst.	# bds	Res meters	Swath Km
IRS-1C,D	WIFS	3	188	810
IRS-P6	AWIFS	3	70	720
SPOT 4,5	Veg.	4	1000	2200
CBERS	WFI	2	260	900
Terra	MODIS	2	250	2330
		5	500	2330
		29	1000	2330

^a Swath is achieved by two side by side instruments
 Multispectral: M, Hyperspectral: H, Panchromatic: P, Radar: R
 VNIR: Visible and Near IR
 SWIR: Short wave IR
 MIR: Mid wave IR
 TIR: Thermal IR

5-year design lives that kept the U.S. land imaging program alive. The radar situation was even more biased as the foreign countries aggressively pursued radar imaging and launched seven radar satellites in this period.

Satellite Types. Before discussing the individual programs, it will be useful to note the data of Table 2. This table divides the current land imaging satellites into four classes; Landsat-like, frequent global coverage, high-resolution, small area coverage, multi & hyperspectral experimental, and radar. The columns provide the ground resolutions of the panchromatic and multispectral bands, the location of the color bands, and the image swath. The similar values of these parameters in each class illustrate the practical results of the trade-offs that must be made among the three principal observation variables, spatial resolution, spectral resolution and band coverage, and temporal resolution or swath width when designing a satellite/sensor system. The table also lists the wide field of view sensors carried by the Indian, French, China/Brazil, and U.S. satellites that provide daily or near daily coverage everywhere at low resolutions, usually 1 kilometer or a bit less. These sensors are similar to the imaging sensors carried on weather and ocean color satellites that are not covered in this article. (See Weather Satellites by Singer and Rao in this *Encyclopedia*.)

The broad swaths of the Landsat class satellites enable them to cover the entire globe many times each year: Landsat and CBERS 23 times; SPOT, when using both sensors, 15; and IRS 18 times a year. The number of times that optical systems can be expected to obtain cloud-free images is one-half to one-third of those figures. The public sector is responsible for mapping and managing large (state to country to global) land areas, so it should not be surprising that all of the satellites in this category are government funded. The high-resolution systems have about one-tenth the swath of the broad area systems, and as a result, they cover the total land cover about twice per year. However, they can point off-track and thus can return to any given site within 1 to 3 days.

The majority of the satellite and sensor data in following sections were taken from Ref. 14. It is not much of an exaggeration to say that this chapter could not have been written were it not for the extraordinary work of H. J. Kramer. Mr. Kramer's 1500 pages recording the engineering and operational characteristics of every sensor and satellite launched into Earth orbit from Sputnik to 2000 will be considered the essential reference volume of Earth and atmospheric observation programs for years to come.

France. The SPOT (Satellite Pour l'Observation de la Terre) series was created by the government's Centre National D'Etudes Spatiales (CNES) and is operated by the SPOT Image Corporation. SPOT Image sells the SPOT images and data products worldwide through a commercial network of distributors and 23 ground receiving stations. The corporation's main shareholders are the government's Centre National D'Etudes Spatiales (CNES—38.5%) and ASTRIUM, the prime contractor for the SPOT satellites (35.6%) (15).

SPOT was designed for the commercial application market from the beginning. Its 10-meter Pan band was pointed to user mapping needs, and its ability to look from side to side made it possible to see any particular site seven times during its 26-day global coverage cycle (compared to Landsat's once in 16 days, see Figure 4). This capability was developed to meet a commonly expressed user requirement to see a site often and at specific times. Four satellites are currently

in operation that give customers the opportunity to see their sites every day. SPOT 1, 2, and 3 had only three color bands. This limited SPOT's value for some types of multispectral analysis of special use in crop monitoring. However, the addition of a fourth band on SPOT 4 and 5 has somewhat mitigated Landsat's advantage in such multispectral applications.

Spot 1 was launched in 1986, 2 years after Landsat was commercialized; the two commercial entities competed fiercely. In 1991, Landsat's sales were \$32 M and SPOT sales were more than \$40 M. Neither figure is close to making the program self-sustaining. Since 1999, SPOT sales have faced serious competition from Landsat 7 because the 1992 law requires selling Landsat data at the "Cost of Filling User Requests" (COFUR) which is presently \$600 a scene, compared to SPOT's roughly \$2000 for a much smaller image. Landsat 7's 15-meter Pan band also cut away much of SPOT's 10-meter advantage. Based on the launch of SPOT 5 (4 May 2002), Spot Image is hoping that its 60-kilometer swath, 2.5 and 5.0-meter resolution images will fill a large mapping niche that neither the new 1-meter, 10-kilometer swath commercial satellites or Landsat 7 can fill. Perhaps even more salable is its currently unique ability to produce continuous stereo strips for the production of high accuracy three-dimensional maps. Spots 4 and 5 also carry the vegetation sensor that the European Union and several of its member countries developed to meet their agriculture monitoring requirements. This sensor provides a daily 1-kilometer global coverage similar to, but with much better spectral data than, the advanced very high resolution radiometer (AVHRR) on POES used by NOAA to produce daily and 10-day vegetation maps of the globe. However, the AVHRR data are provided at the cost of reproduction, as are the data from NASA's moderate-resolution imaging spectroradiometer (MODIS) sensor that has even better spectral data than the vegetation sensor. Thus, the U.S. policy is again in the position of negatively affecting France's commercialization objective.

SPOT plays an important role in the European Community (EC). The Commission of the European Communities is using SPOT to set up an environmental database covering the entire EC. SPOT data are now an integral part of the EC's operational agricultural statistics system used to manage its common agricultural policies.

India. The Indian Space Research Organization (ISRO) manages the Indian Remote Sensing Satellite (IRS) system. The intent of the program is to support India's National Natural Resources Management System (NNRMS). NNRMS supports the national economy in the areas of agriculture, water resources, forestry, and ecology, and the availability of the new high-resolution satellites makes studies of urban sprawl, infrastructure planning, and large-scale mapping possible. The basic use of the IRS data is to support such national infrastructure efforts, but ANTRIX Corp. Ltd., ISRO's marketing arm, sells the data commercially and has franchised sales of its data through the Space Imaging Corp. in the United States (1).

As can be noted in Table 2, IRS-1C and D have four color bands, 23-meter resolution, and a 142-kilometer swath that make it comparable with Landsat in many ways. However, IRS-1C's 5-meter panchromatic capability was the best resolution commercially available on a regular basis until the launch of SPOT 5. (Note that the Russians sold some of their archived intelligence satellite data

scanned to 2 meters resolution in this period). Having launched its Technology Experiment Satellite (TES) in October 2001, India is now the only government orbiting a nonmilitary, 1-meter class mission. Its current plans include launching one more. Its high-resolution systems do not carry a multispectral sensor apparently because color is not required for mapping and tracking civil and military activity/construction. India also plans to match SPOT 5's 2.5-meter relatively broad swath capability with its P-5 satellite.

Japan. Japan's early interest in keeping up with the United States and other space-faring nations began with launcher development and communications satellites. As noted in Ref. 1, "The success of Landsat led to Japanese interest in remote sensing... (and their)... remote sensing strategy therefore followed the so-called "Landsat model"—developing remote sensing satellites for scientific and research purposes." From 1987 to 1996, Japan launched four satellites that had land imaging sensors of better than 50 meters resolution.

The first two, Marine Observation Satellites 1 and 2, launched on 19 Feb. 1987 and 7 Feb. 1990, respectively, focused on ocean and atmospheric data but also carried two land imaging sensors that had 50-meter resolution and 100-kilometer swaths. When combined with a 15-kilometer overlap, their total swath matched Landsat's 185 kilometers exactly. The sensor's 15-kilometer overlap, also provided stereo images. The next satellite, Japan Earth Resources Satellite (JERS-1), launched on 11 Feb. 1992, was devoted totally to Earth resources. It carried both optical and radar sensors; each had 18-meter resolution and a 75-kilometer swath. Its successor, the Advanced Earth Observing Satellite (ADEOS), launched on 17 Aug. 1996, was like the MOS satellites, a combination land, ocean, and atmospheric measurement system. The land sensor, the Advanced Visible and Near-Infrared Radiometer (AVNIR), added a Pan band that had 8-meter resolution and the additional capability of tilting from side to side for faster site-return opportunities. The Japanese have consistently opted for large satellites that have many sensors (Table 2 lists only the land imaging sensors; each had several others). That is often the preferred scientific approach because it provides the opportunity for comparing multiple observations of the same site. The unfortunate early demise of ADEOS has caused a large gap in the Japanese satellite program. They have, however, ambitious plans to continue in the land imaging field and are currently building their Advanced Land Observing Satellite (ALOS) for launch in 2004. ALOS will carry both optical and radar sensors. Its two optical sensors contain advanced features that are required to meet the mission's goals of mapping at 1:25,000 scale without the use of ground control points and with 3.5-meter vertical accuracy for contour plotting. The panchromatic remote-sensing instrument for stereo mapping (PRISM) provides continuous three image stereo (fore, nadir, and aft) and has 2.5-meter resolution and a swath of 35 kilometers. The nadir view has a swath of 70 kilometers. The advanced visible and near-infrared radiometer-2 (AVNIR-2) has four spectral bands at 10-meter resolution and a 70-kilometer swath and has the ability to swing off-track for quick return viewing. Its other major goal of 24-hour response for covering disasters requires both pointing and an all-weather capability of the radar system. It will be an impressive test of the value of a multisensor suite on a single satellite.

Japan has also supplied a major new land imaging sensor, the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) that is currently flying on NASA's Earth Observation Science (EOS) Terra Satellite. See discussion later.

Russia. Russia was the first to use CCD push-broom technology in space for the 18 June 1980 launch of the experimental Meteor-Piroda-5 satellite carrying the MSUE-E (multispectral scanning unit electronic) that had three spectral bands, a spectral resolution of 28 meters, and a swath of 28 kilometers. This successful test was followed up by incorporating several versions of this sensor into a series of land remote sensing satellites similar in function to Landsat, called the Resurs-O1 series. Operated by SRC Planeta, a large conglomerate of scientific and industrial centers, the four satellites provided moderate (45-meter) resolution multispectral images. Resurs -01 #1 was launched on 3 Sept. 1999, #2 on 20 May 1988, #3 on 4 Sept. 1994, and #4 on 7 July 1998. None is currently operational (14). Unlike France, India, and the United States, Russia did not commercialize these data.

However, Russia was the first to offer high-resolution imagery on the commercial market. The Soviet trade association Soyuzkarta offered 5-meter imagery in 1987, and in 1992, two Russian firms began to sell selected images whose resolutions were as low as 2 meters (16). As noted previously, the 2-meter commercial sales were important in pushing the Clinton administration into a commercial licensing policy that has granted licenses for 1- and 0.6-meter commercial resolution systems. The Russian data were digitized from film data taken by their military satellites. The Soviet and the Russian agencies concluded many sales distribution agreements with U.S. and European market organizations (17), but none has been commercially successful. The most aggressive of these programs was SPIN-2 (for Space Information, 2 meter resolution). In July of 1995 Sovinformspутnik signed a contract with a group of American companies, Aerial Images, Central Trading Systems and Lambda Tech International, for a bulk supply of 2-meter imagery of the southeastern part of the U.S. The contract envisioned American customers paying for the launch and operation of a Kometa film carrying satellite. The first attempt on 5/14/96 was lost due to failure of the Soyuz-U rocket. The second launch and flight from 2/17/98 to 5/2/98 was a success. However the commercial sales, in spite of a companion agreement with Microsoft for the development of a system for selling imagery on the Internet, failed to materialize, putting an end to future launches. However, data from a wide range of sensors, including two synthetic aperture radars, are still being offered (17) and are potentially useful to those seeking to chart land surface changes during the last 20 years.

In addition to making images from its military satellites available, Russia has many plans for future satellites (17), but all are apparently on hold at present due to the lack of available resources. However, it can be expected that their considerable experience in optical and radar technology and their ongoing launch and spacecraft activities will someday bring them back into the commercial imaging market.

China. China's first venture into land satellite sensing was carried out in cooperation with Brazil. The China/Brazil (CBERS) satellite carries two multispectral sensors, the HRCC (high-resolution CCD camera) and the IRMSS

(infrared multispectral scanner). The HRCC is a push-broom sensor that has five 20-meter resolution bands and the ability to image off-axis to provide repeat viewing in 3 days. The IRMSS is a whisk-broom scanner like the TM and has the same bands: 76-meter resolution in the reflective bands and 156-meter resolution for the thermal. Both sensors have a 120-kilometer swath. The IRMSS also is like the TM and the ETM in that it has an onboard calibration system that includes both an internal calibrator and a solar calibrator that provide 3% band to band and channel to channel calibration precision and absolute calibration accuracy of less than 10%; these figures compare favorably with the TM. The satellite also carries a wide field imager that has 250-meter resolution and an 885-kilometer swath. The two countries have apparently made no attempt to make the data available in the marketplace. The first CBERS was launched on 10/14/99 and the second is scheduled for 8/10/03. The two countries have also announced plans for CBERS 4 and 5 that will add a 5 meter panchromatic sensor to the multispectral capability of CBERS 1 and 2. China has also launched two satellites of its own, Ziyuan ZY-2A and 2B launched on 9/1/00 and 10/27/02 respectively. Both have panchromatic sensors, the first reportedly of 9 meters resolution and the second 3 meters.

Korea. Korea launched its first satellite on 10 Aug. 1992 as part of its program to gain space technology capability and has supported an increasingly aggressive program since then. KITSAT-3 (Korean Institute of Technology Satellite 3) launched on 26 May 1999 was Korea's first medium-resolution land imaging satellite. Its Multispectral Earth Imaging System (MEIS) sensor has three 15-meter resolution bands and a 50-kilometer swath. The KITSAT series was followed by the KOMPSAT (Korea Multi-Purpose Satellite) series. KOMPSAT-1, launched on 20 Dec. 1999, carries an electro-optical camera (EOC) that has a 6.6-meter Pan band and a 17-kilometer swath. It is designed to obtain cartographic maps of Korea at 1:25,000 scale. It also carries an ocean color sensor. KOMPSAT 2 is being developed for launch in 2004 to provide surveillance of large-scale disasters. The plan is for it to carry a 1-meter Pan and a 4-meter, four-band multispectral sensor (14).

The Current Period

The third period, 1997 to the present, is dominated by the entrance of commercial high-resolution systems, but includes the return of U.S. activity in the form of an R&D program and the launch of the long-awaited Landsat 7. In addition, as discussed before, Korea, China, and Brazil, the latter two in a cooperative program, joined the civil land imaging satellite community in this period.

Commercial Satellites

As can be seen in Table 2, high-resolution systems image objects that range in size from 1.8 to 0.6 meters (6 to 2 feet). These resolutions are about the same as the original classified systems of the early 1960s. The private sector dominates in this capability range. Three of the four high-resolution mission satellites launched successfully to date have been commercial systems.

U.S. commercial systems are the direct result of the 1992 Land Remote Sensing Policy Act and the subsequent 1994 Clinton administration policy that led to licensing commercial resolutions as high as 0.6 meters. The commercial sector, after ineffectually pursuing a market in Landsat and SPOT images is hoping that satellite imagery of 1 meter or less resolution can be profitably marketed at prices that will capture some of the very large aircraft remote sensing market and create entirely new markets based on the satellite's ability to gather scenes anywhere, anytime, for anyone, clouds willing. The commercial companies are also well aware that their 0.6 to 1.0-meter resolutions are sharp enough to provide very useful information to the military and are pursuing both national and international defense markets.

These high resolutions have been achieved by restricting the width of their images to between eight and 16 kilometers and for those with color bands, limiting the spectral coverage to the VNIR bands. The narrow swath does restrict their utility for the broad area monitoring tasks of the Landsat type systems, but the satellites' ability to point off-track provides the capability to image any specific target at 1 to 2-day intervals. This agile pointing enables the acquisition of stereo images and thus the production of three-dimensional images.

Three American companies, Space Imaging (SI), Digital Globe (DG) (formerly Earth Watch), and OrbView and one Israeli company, ImageSat International (ISI), are currently offering high-resolution images commercially. They have had their share of launch failures. Prior to the successful launches, Digital Globe lost its first two satellites at or shortly after launch; OrbView, SI, and ISI, their first. Launch insurance has enabled them to survive to date. All recognize that their business cannot survive with single satellites, and the three U.S. companies are each planning follow-up satellites. ISI has published a very aggressive plan for six more launches by 2006 of a new and improved system that essentially copies Digital Globe's half-meter panchromatic and four-band multispectral capabilities. The timing of all of these future systems is somewhat uncertain because it depends on each company's ability to obtain the necessary funding from their slowly growing sales or from the currently troubled financial marketplace.

The technical success of the commercial systems has proven that the private sector can meet the technological challenges in providing sophisticated data products. As each company struggles to turn its technical success into a financial success, it is faced with a variety of challenges. The very technologies that have made spacecraft imaged data possible, digital sensors and GPS (the satellite Global Positioning System), have also increased the capabilities of their aircraft-based competition. The satellite's ability to get images of any and all countries is their prime advantage over their aircraft-based competitors. That feature is of interest to many multinational companies and to some nongovernmental organizations (NGOs). It is of great interest to governments for military and political reasons. The U.S. companies are pursuing agreements with foreign countries, but they face the perception that the "shutter control" clause in their government license would subject the availability of data to the whims of U.S. foreign policy. Several foreign governments have already announced plans for their own high-resolution satellites (see The Future section). In the United States, the administration is well aware of military utility of high-resolution data and has instructed NIMA (the

National Image and Mapping Agency) to use commercial systems to expand its capabilities and to reduce the requirements on its classified systems.

There may be another market niche. As noted previously, the Congressionally mandated commercial data buy for the Landsat Data Continuity Mission has prompted two companies, Digital Globe and Resource 21, to compete to provide NASA with Landsat data as a by-product of their own projected satellite data service, weekly crop health reports to farmers. These reports will enable farmers to take early remedial actions on the less productive areas, the so-called "precision farming" approach made possible in part by another satellite service, GPS. These systems will have the coverage and spectral capabilities of Landsat but at higher resolutions, 7.5 or 10 meters and more frequent coverage, 4 or 8 days.

At this time, it must be concluded that the commercial land imaging satellite programs are still in their start-up stage and that their future is still uncertain. The interested reader is advised to seek out their web pages to keep up with their progress.

Multi- and Hyperspectral Tests

These systems, 3 funded by the U.S. government and one by ESA, are space tests of advanced sensors. As shown in Fig. 6, the sensors include more and narrower bands to provide increased sensitivity for analysis. Aircraft tests have demonstrated that measurements of the radiance of essentially all of the spectrum using the so-called hyperspectral sensor provides increased ability to identify a wide range of geologic material and crops as well as to discern crop health and potential yield. Two of the new sensors have added increased thermal coverage that has also been demonstrated to improve crop classification and health status.

ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer), a Japanese sensor flown on NASA's Terra mission, has twice the number of Landsat's 7's spectral bands. It is actually a set of three telescopes flown together. The VNIR system includes a backward pointing telescope that is combined with its nadir telescope to provide 15-meter resolution, continuous stereo data. The six shortwave IR (SWIR) bands and the five thermal IR (TIR) bands, all at 15-meters resolution, are opening new analytical capabilities.

The MTI (Multispectral Thermal Imager) was developed by the DOE's Office of Nuclear Non-Proliferation to test the ability of multiband thermal data to measure the heat output of nuclear plant-cooling and thereby its power level. In addition to its four standard VIS 5-meter bands, it includes a unique series of 20-meter bands, three NIR, three SWIR, two MWIR, and three TIR. The DOE 3-year program includes imaging industrial, government, and natural sites. Unfortunately, its data have not yet received the attention of the broader research community they deserve.

EO-1: This NASA spacecraft is designed to test advanced technologies of potential use in the next Earth resources satellite. It carries two 30-meter resolution sensors: the Advanced Land Imager (ALI) with six VNIR bands, three SWIR bands, and one 10-meter resolution pan band; and a hyperspectral imager, Hyperion, which has 220 bands covering the VNIR and SWIR range. It has been placed in orbit near Landsat 7 to obtain comparative data sets. An extensive scientific and

application analysis program is in process as this is written. It has proven so useful that EO-1 operations are being continued beyond their planned duration by selling data to several government agencies. The ALI push-broom approach resulted in a sensor that is one-quarter the weight and one-seventh the volume of the ETM+ but provides three more bands and a 10-meter pan band (18).

PROBA: PROBA is a technology demonstration satellite designed to demonstrate many advanced satellite technologies in a manner similar to EO-1. It carries CHRIS, an experimental hyperspectral imager that can be programmed to produce either 19 VNIR bands with 18 meter resolution or 63 bands with 36 meter resolution. It also carries an 8 meter resolution panchromatic sensor.

Radar

Radar images through clouds and at night; this can be very important in those parts of the world like the tropics and Europe that are cloud-covered a high percentage of the time. Their sensitivity to a target's size, shape, and material characteristics and to its moisture content provides an entirely different information set than an optical system's measurement of reflected and emitted radiation. Radar sensors come in a variety of types based on their signal frequency and signal and reception polarization. To date, all satellite radar systems have been single frequency, which, like panchromatic optical data, are very valuable for mapping type functions but are not very good at classification functions.

Radar has not yet been used in civil Earth observation satellites to the extent that the optical sensors have. As the discussion on future systems below notes, this may change in the next decade. Figure 1 shows the orbit history of the civil radar satellites to date. Seasat was a multisensor satellite developed by NASA in cooperation with the Navy. Its SAR sensor is regarded as the first imaging SAR system used in Earth orbit (14). Launched 27 June 1978, it failed abruptly 106 days after. In spite of its short life, it was a great success. The quality of its data was responsible for ESA's 1981 decision to favor development of a radar system. ESA was stimulated by radar's ability to image through Europe's frequent cloud cover (19). As can be seen in the figure, ESA has maintained the continuity of its radar program to the present. In the United States, radar data were apparently judged so potentially valuable for military use that civil radar satellites were discouraged, and there have been no U.S. civil radar satellites to date. The current Defense Department position limits civil systems to five-meter resolution, a restriction that caused Canada to cancel their contract with an American firm and turn to Italy to build its planned 3-meter system.

At present, the uses of single-frequency systems are many and impressive. A short list would include oil spill detection, ice movement monitoring, mapping (one of Canada's prime applications), monitoring forests and agricultural land in the tropics, geological mapping aided by its oblique viewing angle and vegetation penetration capabilities, and disaster evaluation, especially cloud-covered floods. Using data from two overflights, radar satellites can provide interferometric maps of surface elevation changes of the order of a few centimeters.

Current and planned radar satellites come in two styles. ERS-2, Envisat, and ALOS-1 carry many other observation sensors along with their SAR. Canada's current and planned Radarsats are dedicated to their SAR mission.

A discussion of radar satellites would not be complete without mentioning the NASA-sponsored Shuttle-based radar sensor program. Although restricted from creating free-flying radar satellites, NASA's Jet Propulsion Laboratory (JPL) was able to advance the technology in a series of five Shuttle missions carrying radar sensors. SIR A, flown in 1981, and SIR B, flown in 1984, carried L-band sensors. The third and fourth Shuttle missions, SIR C/X-SAR, flown twice in 1994 in cooperation with Germany and Italy, consisted of separate but connected L, C, and X-band SARs. The long and complex analysis program of that data has shown that multifrequency radar analysis can serve many of the applications that multispectral analysis has made possible for optical systems, which, in effect, could lead radar into closer competition with the now dominant optical world. The fifth Shuttle mission produced what may be space-based radar's most useful accomplishment to date. The 10-day Shuttle Radar Topographic Mission (SRTM) was sponsored by NASA and NIMA and launched on 11 February 2000. It carried two SIR-C/X SARs separated by a 200-ft boom attached to the Shuttle's bay. The data from the resulting continuous stereo view produced a topographic map of the land surface on which 95% of Earth's population lives and which is 80% of the total global land surface. Surface elevation of all ground points 30 meters apart will be provided to scientific investigators, and elevations 90 meters apart will be open to the public.

The Future

Predicting satellite launch dates, even by those who are paying for them, is a risky business. Since 1995, the author has tracked the progress of more than 30 land imaging satellites. The delays between the first announced launch and the actual launch date averaged a bit more than 2 years, and nine of the 30 had an average of a 4-year delay. Keep this history in mind while reading the following discussion.

Even without firm knowledge of the status of many of the satellites being proposed, one thing is certain, that civil land imaging satellites will be spread over the globe. This is dramatically illustrated by Fig. 15. As its caption aptly states, "Global competition is heating up." There are 16 foreign countries that have plans for land imaging satellites. The 1-meter and better satellites, currently mostly the province of the U.S. commercial sector, will be seeing a significant part of this global competition in the form of government systems in France, Italy, Korea, and India, as well as the multisatellite plans of Israel-based ISI.

The most mature of these plans have been noted in the previous sections, but as this is being written, both governments and private corporations are discussing/announcing ever more ambitious plans. Rapid/Eye of Germany continues to promote its plans for a commercial four-satellite, 6-meter system to provide a crop monitoring service similar to that being proposed by the LDCM contractors. The French and Italians are discussing the development of a five-satellite, 1-meter satellite system to be launched by 2005/2006 made up of the French Pleiades two-satellite optical system and the Italian Cosmos-Skyimed three-satellite radar program (14). The German

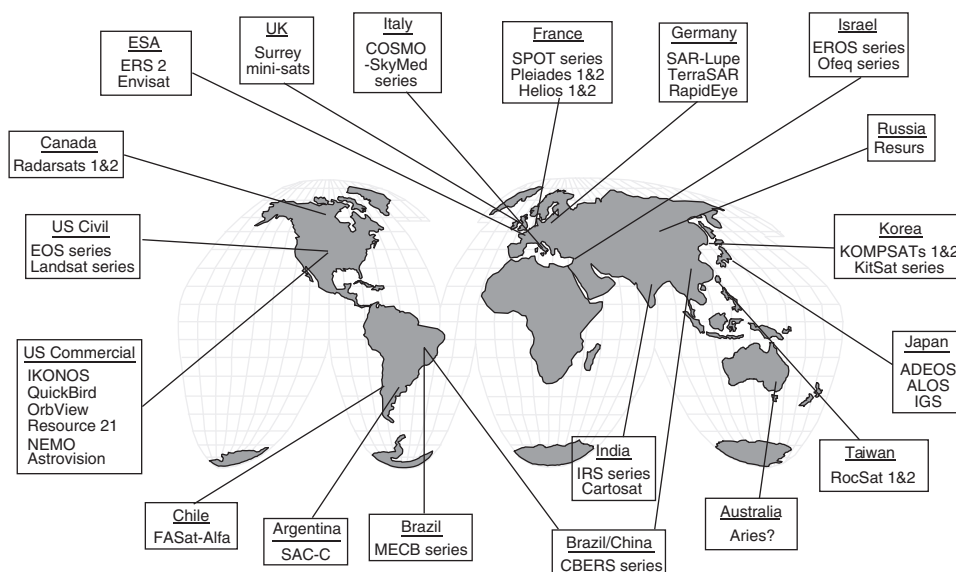


Figure 15. Global competition is heating up. Reproduced by the kind permission of John Baker and the Rand Corporation.

government announced TerraSarX and TerraSarL, a 1-meter radar satellite program planned for launch in 2006. These somewhat competing programs are the subjects of much debate in the European Union planning process. However, should most of these plans carry through to launch, there could be 19 1-meter or better systems in orbit by the end of 2006, including three to five radar systems.

A new approach to combining daily coverage and medium resolution land observation capability by creating a constellation of small low cost microsatellites (as much as 1/5th the cost of current commercial systems) is being developed by the British company, Surrey Satellite Technology Ltd. (SSTL) (20). An early example is already in orbit. UoSAT-12, developed and funded by SSTL, carried payloads from Singapore and ESA into a 64° inclination orbit on a Russian launcher on 12 May 1999. Its payload included one 10-meter resolution panchromatic sensor that had a 10-kilometer swath, one 975-meter resolution sensor that had a 1000-kilometer swath, and two 32.5-meter resolution multi-spectral sensors that had 33-kilometer swaths. SSTL is currently building a five-satellite Disaster Monitoring Constellation (DMC) that will be capable of providing 32-meter resolution, three-band color images of any location once per day (14). The first mission sponsored by Algeria was launched from Russia on November 28, 2002, and four more are scheduled in 2003 (21). The other sponsors are Nigeria, UK, Thailand, and Turkey. China and Vietnam are planning similar systems for launch in 2004 that carry a 4 meter panchromatic sensor. It is clear from the number of nations involved, that remote sensing of one's land from space brings out a strong desire to have one's own satellite for reasons that appear to be a mixture of security, practical applications, and simple national pride.

Summary

Of all efforts to date to use our ability to operate satellites in space for the betterment of life on Earth, the U.S. Landsat program certainly ranks among the most successful. It invented and implemented a whole new technology, a truly new and powerful way of looking at Earth through the analysis of its reflective and emitted radiation, a method that for the first time has made it possible and practical to monitor the entire surface of Earth on a regular basis at resolutions capable of observing and assessing human-scale activities. At the same time, the Landsat program and other government programs actively spread the data and the technologies involved throughout the whole world, with specific emphasis on third-world countries. The technical achievements are significant, but it may be that this worldwide sharing will be its enduring monument.

All of this was accomplished despite the difficulties the program encountered from before its birth to the present, difficulties related to national security issues, agency roles, delays in data delivery, funding uncertainties, and a failed attempt to commercialize the federal program. Its 30 years of synoptic global data remain the only systematic collection of consistent global land images. The current U.S. policies of having Landsat image the entire global land surface four times a year and of providing data to all for the cost of reproduction has yet to be emulated by any of the foreign satellite systems.

The rest of the world started their civil land imaging programs 15 years later, and they have launched more than twice the number of missions. France and India equaled or bettered the launch pace of the early Landsat years, and both countries have strong plans to continue and improve their systems. Japan and Russia started just as vigorously, but halted after their late 1990 launches. Japan has seemingly firm plans to restart its program in 2004, and Russia's future plans remain uncertain. Although foreign optical satellite programs have been very impressive and in some ways more innovative than the U.S. program, their image acquisition programs have focused on getting data for sale in the marketplace rather than on creating a synoptic archive of global images for scientific use.

The radar history is entirely different. The U.S. short-lived Seasat radar mission provided sufficient data for Russia, Japan, ESA, and Canada to create radar satellite programs that, Russia excepted, have been maintained to the present. To date, all of the civil land imaging radar satellites have been foreign, and current planning statements indicate that this situation will continue into the foreseeable future. This is due to a combination of reasons, including Europe's persistent cloud cover, Canada's strong interest in mapping the changing ice fields of their North, and the reluctance of the U.S. defense establishment to allow civil radar systems.

Until the late 90's all of the systems were funded by their respective governments. As a result of the U.S. 1992 Land Remote Sensing Policy Act that permitted the development and operation of civil satellites with resolutions in the meter or less range, the pull of the marketplace entered the land imaging satellite game. This was a significant break with the tacit international policy of keeping high-resolution "spy quality" data as the exclusive province of

governments. This opening of high-resolution data to the civil sector was helped by Russia's commercial sale of 2-meter data, but was vigorously protested by France as well as by U.S. defense agencies. Congress' motivation was based on the belief that meter-level systems could compete with aircraft data for that multibillion dollar market and that sharing the cost of such satellites with a commercial market would dramatically reduce the government's data costs. At this time, it still is uncertain if purely commercial systems will find a sufficient market share to ensure their future.

In any case, the secrecy genie is out of the bottle. The availability of meter-scale data in the open marketplace has opened the door to data uses far beyond the scientific and large-scale land management functions of the Landsat resolution data. The door has been opened to global information transparency that can change the relationships between nation and nation and between citizens and their governments (1). The political, scientific, and commercial currents during the next 25 years of Earth-observing systems will be no easier to chart than were the first, but the systems they spawn will certainly advance the understanding and informed use of the planet's resources and aid in establishing the trust among nations, that will make global environmental management possible.

Further Reading

The history, description, technologies, data analysis and data distribution of the Landsat program up to 1997 are covered in detail in the journal *Photogrammetric Engineering & Remote Sensing* 63 (7): July 1997 that commemorated the twenty-fifth anniversary of Landsat. The authors are individuals closely involved with that aspect of Landsat activity they treat, and the papers listed must be considered as basic source material:

Lauer, D.T., S.A. Morain, and V.V. Salomonsons. The Landsat program: Its origins, evolution, and impacts.

Mika, A.M. Three decades of Landsat instruments.

Thome, K., B. Markham, J. Barker, P. Slater, and S. Biggar. Radiometric calibration of Landsat.

Landgrebe, D. The revolution of Landsat data analysis.

Drager, W.C., T.M. Holm, D.T. Lauer, and R.J. Thompson. The availability of Landsat data: Past, present, and future.

Williamson, R.A. The Landsat legacy: Remote sensing policy and the development of commercial remote sensing.

Goward, S.N., and D.L. Williams. Landsat and Earth systems science: Development of terrestrial monitoring.

Unger, S.G. Technologies for future Landsat missions.

Colvocaresses, A.P. Landsat application to nautical charting.

Several of the references are broad sources of information and deserve detailed reading by anyone interested in this subject matter.

The following two works provide detailed histories of Landsat development and interesting analysis of the whys and wherefores of that tortuous process:

Mack, P.E. *Viewing the Earth: The Social Construction of the Landsat Satellite System*. MIT Press, Cambridge, 1990.

Thomas, G.B. *Analyzing Environmental Policy Change: U.S. Landsat Policy, 1964–1998*, Doctoral Dissertation, Colorado State University, Fort Collins, CO, 1998.

The history and current global status of civil land imaging systems are provided by the papers written by international authors on experiences and plans of their own countries.

Baker, J.C., K.M. O'Connell, and R.A. Williamson (eds). *Commercial Observation Satellites: At the Leading Edge of Global Transparency*.

The penultimate source of all technical information about the sensors and satellites that have and are about to observe Earth from space is the following monumental work. In addition to the technical details, it also contains excellent histories of Landsat and the other major development programs.

Kramer, H.J. *Observation of the Earth and Its Environment: Survey of Missions and Sensors*, 4th ed. Springer, 2002.

This paper provides a detailed history of the events preceeding and following the '92 Landsat law and the creation of Landsat 7.

Sheffner, E.J. The Landsat program: Recent history and prospects. PE&RS 735–734.

During the Landsat development years, NASA and other agencies called upon the National Academy and its various committees to provide analysis and advice. The following are some of the resulting reports and publications.

Remote Sensing of the Earth from Space: A Program in Crisis. Space Applications Board, National Academy Press, 1985.

Earth Observations from Space: History, Promise, and Reality. Space Studies Board, National Research Council, 1995.

Liverman D., Moran, E. F., Rindfuss, R. R., Stern, P. C. (eds). *People and Pixels: Linking Remote Sensing and Social Science*. National Academy Press, 1998.

Ecological Indicators for the Nation. National Academy Press, 2000.

A useful examination of the policy issues:

Harris, R. *Earth Observation Data Policy*. Wiley, 1997.

A basic textbook on remote sensing and image interpretation:

Lillesand, T. M., and R. W. Kiefer. *Remote Sensing and Image Interpretation*, 3rd ed. Wiley, 1994.

A detailed guide to the theory and practice of multispectral imagery:

Multispectral Imagery Reference Guide. Logicon Geodynamics Inc. Spectral Imagery Training Center, 1997.

The Internet

The above suggested reading will be useful. However, the field is a rapidly evolving one and the reader is well advised to make use of the internet. The NASA and USGS home sites have directions to sites specifically devoted to Landsat and other satellite imagery. All of the commercial companies have their own sites, easily located by any search engine. The same is true of all of the countries involved in satellite sensing. Finally, news about the status of any particular satellite can often be found by typing its name into Google.

BIBLIOGRAPHY

1. Baker, J.C., K.M. O'Connell, and R.A. Williamson (eds). *Commercial Observation Satellites: At the Leading Edge of Global Transparency* 2001.
2. Hall, S.S. *Mapping the Next Millennium: How Computer-Driven Cartography Is Revolutionizing the Face of Science* 1992.
3. Mark, H. *The Space Station: A Personal Journey*. Duke University Press, Durham, NC, 1987.
4. Burrows, W.E. *Deep Black, Space Espionage and National Security*. Random House, New York, 1986, p. 401.
5. Lauer, D.T., S.A. Morain, and V.V. Salomonson. The Landsat program: Its origins, evolution, and impacts. *Photogrammetric Eng. Remote Sensing* 63 (7): 831–838 (1997).
6. Mack, P.E. *Viewing the Earth: The Social Construction of the Landsat Satellite System*. MIT Press, Cambridge, 1990.
7. Lowman, P.D. Jr. Landsat and Apollo: The forgotten legacy. *Photogrammetric Eng. Remote Sensing* 65 (10): 1143–1147.
8. Pecora, W.T. Remote sensing of Earth resources: Users, prospects and plans, NASA's long-range Earth resources survey program. Thirteenth Meeting, Panel on Science and Technology, Committee on Science and Astronautics, U. S. House of Representatives, 25 January 1972. U.S. Government Printing Office, Washington D.C.
9. Drager, W.C., T.M. Holm, D.T. Lauer, and R.J. Thompson. The availability of Landsat data: Past, present, and future. *Photogrammetric Eng. Remote Sensing* 63 (7): 869–875 (1997).
10. UN/OOSA, 1994, United Nations Treaties and Principles on Outer Space, Office for Space Affairs, UN Office at Vienna. Document A/AC. 105/572 pp. 43–46.
11. Thomas, G.B. Analyzing environmental policy change: U.S. Landsat Policy, 1964–1998, Doctoral Dissertation, Colorado State University, Fort Collins, CO, 1998.

12. Hovis, W., J. Nichols, A. Potter, F. Thomson, Compilers, J. Harnage, and D. Landgrebe (eds). Landsat-D Thematic Mapper Technical Working Group Final Report, NASA, Johnson Space Center, JSC-09797, June 1975.
13. Brown, G.E. Earth observations and global change decision making: A national partnership. Remarks to the National Press Club, October 23, 1991.
14. Kramer, H.J. *Observation of the Earth and Its Environment: Survey of Missions and Sensors*, 4th ed. Springer, 2002.
15. Soubes-Verger, I., and X. Pasco. In J.C. Baker, K.M. O'Connell, and R.A. Williamson (eds), *Commercial Observation Satellites: At the Leading Edge of Global Transparency*, 2001, pp. 187–204.
16. Taha, G. Russian remote sensing programs and policies from commercial observation satellites in Ref. 1.
17. Taha, G.J. in Ref. 1, pp. 165–185.
18. Unger, S.G. Technologies for future Landsat missions. *Photogrammetric Eng. Remote Sensing* 63 (7): 901–905 (1997).
19. European Space Agency. 20 Years of European Cooperation In Space '64–'84. ESA Report 1984, ESTEC, 1984.
20. Florini in Ref. 1, p. 445.
21. Personal communication from SSTL CEO.

WILLIAM E. STONEY
Mitretek Corporation
Reston, Virginia

COMETS

Introduction

The solar system formed 4.6 billion years ago from a cloud of dust and gas called a molecular cloud (1). The molecular cloud was composed predominantly of hydrogen and helium gases and a small amount of heavier atoms and molecules that had been formed during the lifetimes of previous generations of stars. In addition, refractory (heat resisting) dust grains were also present in the cloud. Something, perhaps a nearby supernova, caused increased density regions to form within this molecular cloud. If these cloudlets became massive enough, then they would gravitationally attract the nearby gas, causing the cloudlets to increase in mass and contract. Any residual motion or small rotation within the gas would cause these cloudlets to spin more rapidly as they contracted under gravity. These spinning cloudlets formed the center of the protosolar nebula. The cloud collapsed from the inside out. The rotation caused the nebula to flatten into a disk with most of the mass in the disk and in a bulge at the center. However, there was still some gas in a halo around the disk.

As the central bulge collapsed, it added more material from the surrounding disk and halo. Gravitationally, the center collapsed to a smaller sphere, and the pressure and temperature increased. When the pressure and temperature were sufficiently large (temperature of a few million degrees), the hydrogen atoms in the core of the sphere started to fuse into helium, and the cloud became a new

star. At this point, the star was still surrounded by a disk of material and a halo of gas. The high luminosity of the infant star propelled a “wind” of material flowing outward along its rotational axis, creating a bipolar flow. This blew away much of the material in the surrounding halo, stopping the infall onto the disk. It is within this disk that the planets themselves formed. We can estimate the minimum mass that this disk must have been to form the current planets by several methods: the percentage of rocky material in the gas (0.4% of the total) or the mass-loss rates of young stars ($3 \times 10^{-8} M_{\odot}$ /year for 3 million years or a total mass loss of $0.06 M_{\odot}$ where M_{\odot} is the current mass of the Sun). Thus, the disk from which the planets formed must have had a substantial fraction of the mass of the Sun and much of it was lost after the planets formed.

About this time, clumps of rock and ice started to grow in the disk. The clumps began to collide with one another and accrete (stick together), causing them to grow larger. Larger objects have greater gravitational attraction, so the biggest clumps grew bigger and “gobbled up” the surrounding material. The clumps grew until a critical size was reached, and then they collapsed into the cores of the planets. The largest core became our current-day Jupiter and was 20–30 times the mass of Earth after a few million years (it is currently 300 times the mass of Earth). Around the time the cores achieved such large masses, they began to have profound influences on the surrounding leftover material.

The timescales for these processes are short: the cloud collapse took around 10,000 years, the disk/wind clearing took around 100,000 years, and the initial building of planets was complete in around 10 million years (2).

As the Sun settled into the main part of its life as a normal star, the gravitational attraction from the protoplanets, along with the wind from the Sun, gradually removed the residual material from the disk. Some time after the molecular cloud started to collapse, our solar system consisted of the Sun, the major planets, and small quantities of leftover building blocks of planets, called planetesimals. The nature of these leftover planetesimals depended on where they were in the solar system relative to the Sun. Those nearer the Sun, in the region from Mercury through the asteroid belt (the so-called terrestrial planet zone), were predominantly rocky bodies. In the region from Jupiter onward (the so-called giant planet zone), the bodies were a combination of ices and rocky material. Note that “ice” in this context does not refer only to water ice, but includes ices of other more volatile materials such as carbon dioxide, ammonia, methane, and formaldehyde. The leftover planetesimals from this early period that still reside in our solar system (more on their fate later) are called comets. The comets of today are mostly unchanged from the time they were formed. *Thus, comets today preserve information on conditions when the solar system was formed and are of prime importance for providing constraints on conditions in the early solar nebula.*

Physical Properties

Comets present a striking appearance in the night sky. Their presence has been noted since ancient times; their name is Greek meaning “hairy ones.” Our modern understanding of their nature dates to the 1950s (3). Basically, a comet is a

“dirty snowball.” It is a small body; typical comets have nuclear dimensions of 1–20 km. This nucleus is composed of ices and dirt. As the comet approaches the Sun in its orbit, the nucleus is heated, causing the ices to sublime (sublimation is the process whereby materials go from the ice stage to the gas stage directly without ever becoming liquid). These gases form a coma around the nucleus as they flow outward from the nucleus into the vacuum of space. As the ices sublime and flow outward, they carry dust with them into the coma. The coma of a comet is not a bound atmosphere because the gravitational pull of the tiny nucleus is small. Instead, the coma gases are constantly being replenished via sublimation, and the gases in the outer coma are being lost into interplanetary space.

Once the coma has formed, it is extremely difficult to image the nucleus of the comet with ground-based or Earth-orbital telescopes because it is shrouded by the coma. Only two comets have had their nuclei fully resolved by imaging. Comet Halley was imaged by the Giotto and Vega spacecrafts in 1986. Using the high-quality Giotto images, we measured the nucleus as $8 \times 8 \times 16$ km, but it was actually quite an irregular shape. Comet Borrelly was imaged by the Deep Space 1 spacecraft with a size of $4 \times 4 \times 8$ km and, again, was quite irregular. Other comets have had nucleus sizes estimated by various other techniques. Unfortunately, the spacecraft did not travel close enough to Halley or Borrelly for the comet’s gravity to influence the spacecraft orbits, so we do not have any direct measurement of the density or mass of a comet. Various assumptions have led to the concept that the density is between 0.2 and 1.2 g cm^{-3} (water has a density of 1.0 g cm^{-3}).

As the comet continues in its orbit around the Sun, the Sun is exerting an outward force known as radiative pressure on the dust, pushing the small dust grains away from the Sun. Thus, the dust ends up lagging the comet in its orbit around the Sun. The combination of the radiative pressure of the Sun and the Keplerian motion of these particles as they continue in orbit around the Sun causes the dust to curve away from the comet on a track outside of the cometary orbit and form a tail known as a dust tail. However, not all comets develop dust tails.

Interplanetary space is not empty, and the comet interacts with its local environment. In addition to emitting radiation at most wavelengths, the Sun sheds ionized material in the form of the solar wind. The solar wind is a hot ($\sim 100,000$ K) plasma consisting of an electrically neutral mixture of ions (mostly electrons and protons) that originates in the solar corona, is accelerated into interplanetary space, and drags along the solar magnetic field. By Earth’s orbit, the bulk velocity of the solar wind reaches a maximum of $200\text{--}800 \text{ km sec}^{-1}$. Comets have no intrinsic magnetic fields or ionospheres, but they are continually releasing neutral gases through sublimation. Some of these gases become photoionized. Some tens of thousands of kilometers on the sunward side of the nucleus, a bow shock forms (a sharp boundary separating the ionized material of the comet, called the magnetosphere, from the solar wind material which is forced to flow around the comet) at the distance where the kinetic pressure of the solar wind particles equals the magnetic pressure of the ionized gases. As the solar wind magnetic field drapes around the comet, the solar wind exerts a tangential force on the boundary on the side away from the Sun, dragging the plasma into a long tail called the magnetotail. On opposite sides of the

magnetotail, the current travels in opposite directions. Inside the magnetotail, where the opposing currents meet, is a neutral region called the neutral current sheet. In addition, the solar wind reacts with the ionized coma gases, pushing the ionized gases outward. This forms what is known as the ion tail of a comet. The ion tail is straight and always points directly away from the Sun, which means that it sometimes leads the comet in its orbit. Not all comets display ion tails.

Figure 1 shows a picture of comet Hale–Bopp. This comet was discovered in 1995 by Alan Hale and Thomas Bopp (independently) and came closest to the Sun on 1 April 1997. This image was obtained with a CCD camera (an electronic detector) and special filters on the 0.76-m telescope of McDonald Observatory. The various parts of the comet are labeled in this picture.

The nucleus is quite small, but the dimensions of the rest of the comet are substantial. A typical cometary coma is more than 1 million km in radius, or substantially *bigger* than the Sun. However, as large as the coma is, it is quite insubstantial and has a density that qualifies it as a better vacuum than can be created in most laboratories (a pressure of $10^{-8} - 10^{-9}$ torr)! The tail of the comet can stretch to several hundred million km behind the comet, but the tail has densities even lower than the coma.

A comet does not shine from its own radiation. Instead, the light from the Sun is reflected from the dusty material, and the gas generally “glows” by a process known as fluorescence. (Fluorescence occurs when an atom or molecule reaches an excited state by absorbing a photon and then returns to a lower state by emitting a photon. Fluorescence is termed resonance fluorescence when the wavelength of the

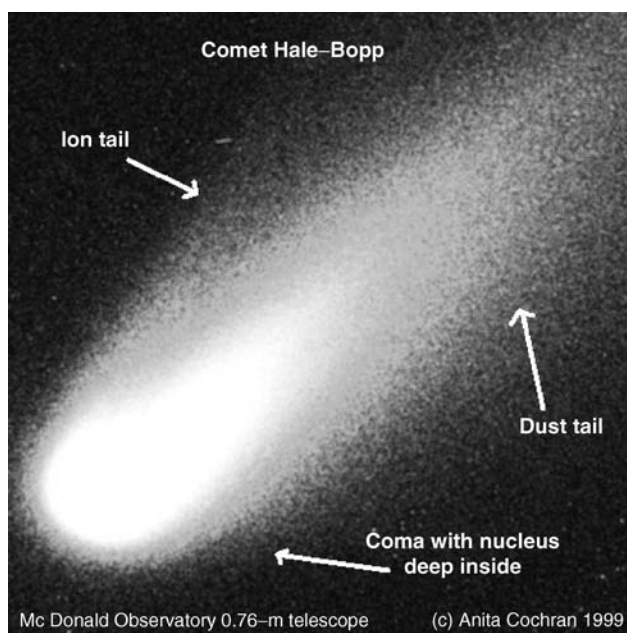


Figure 1. An image of comet Hale–Bopp obtained by the 0.76-m telescope of McDonald Observatory. The image is approximately 40 arcmin on a side. The parts of the comet are labeled on the picture. This is an image obtained by a CCD camera.

exciting photon has the same energy as that of the downward transition and the atom or molecule goes back to the energy level from which it started).

Very far from the Sun, there is not enough heat to sublime the gases of the comet. Therefore, if a comet were viewed at large distances from the Sun, it would appear just as a small, rocky, icy body that has no coma. However, comae have been detected around cometary nuclei at distances past Saturn's orbit [e.g., (4)].

Cometary Orbits and Origins

A comet is generally discovered as a fuzzy object that is moving with respect to the background stars. About half of the comets are discovered by amateurs, who scan the skies looking specifically for comets, and the other half are discovered by professional astronomers who are obtaining images of the sky when they discover the comet. Discoveries of comets are relayed to the International Astronomical Union's (IAU) Central Bureau for Astronomical Telegrams for confirmation and assignment of a designation.

Each cometary discovery is given an official designation by the IAU that consists of the year of discovery followed by a letter identifying in which half month of the year the discovery is made (e.g., January 1–15 is A, January 16–31 is B, ... , December 16–31 is Y; the letter I is not used for half month designations) and by a consecutive number to indicate the order of discovery in that half month. Thus, the fourth comet discovered in the first half of March 1996 would have been designated 1996 E4. This gives each comet a unique name. In addition, comets are named for their discoverer(s); up to three names are assigned to the comet. Thus, the comet that is shown in Fig. 1 is officially comet 1995 O1 (Hale–Bopp) and was the first comet discovered in the second half of July 1995. It was discovered by Alan Hale and Thomas Bopp.

The position of the comet needs to be measured carefully with respect to a fixed reference frame of distant stars or quasars. These positions are then used to compute an orbit around the Sun, taking into account the gravitational influence of all of the planets. All known cometary orbits are elliptical; the eccentricity e of the orbit ranges from close to circular, in the case of comet Schwassmann–Wachmann 1, to parabolic. (Note that the orbits of the planets are very close to circular, except for Pluto.) If it takes the comet fewer than 200 years to orbit the Sun, a comet is termed to be in a periodic, or short-period, orbit. Strictly speaking, all comets are in periodic orbits, in the sense that they are gravitationally bound to the Sun, but some of the comets have periods that are so long compared with a human lifetime that they are designated as long-period or nonperiodic comets. The division at 200 years is purely arbitrary and historical.

The orbits of the major planets are all confined, more or less, to a single plane. If we define the plane of the ecliptic as the plane of Earth's orbit around the Sun, we find that, except for Pluto, all of the major planets are in orbits that are inclined by no more than 7° from the plane of the ecliptic. Even Pluto is inclined only 17° from the plane of the ecliptic. All of the planets also travel around the Sun in the same direction, which we call the prograde or direct sense. Prograde orbits confined near the plane of the ecliptic are consistent with our picture of the formation of the solar system from a disk of material surrounding the newly

forming Sun. For convenience in describing the location of objects in our solar system, astronomers use a distance known as an *Astronomical Unit* or AU, which is defined as the mean distance of Earth from the Sun. It is approximately 1.5×10^8 km. The heliocentric distance of an object is its distance from the Sun.

Cometary orbits have properties very different from the orbits of the planets. Taken in aggregate, there is no preferred inclination of cometary orbits to the plane of the ecliptic. Some comets have orbits that travel around the Sun in the direction opposite from the planets, or retrograde. An example of a comet that has a retrograde orbit is comet Halley.

For now, we will ignore the short-period comets and concentrate on the orbits of the comets that have long- or non-periodic orbits. The semimajor axis of an ellipse is designated a . The inverse $1/a$ of the semimajor axis is proportional to the orbital binding energy, which is the additional energy a comet would need to escape the gravitational attraction of the Sun completely. If $1/a < 0$, the comet is not bound to the Sun.

For long-period comets, we can compute a quantity $1/a_0$, the inverse semimajor axis of the *original* orbit, where the original orbit is the orbit of the comet integrated backward in time to the point before it entered the planetary system and referenced to the barycenter of the solar system. Figure 2 is a histogram of $1/a_0$ (data from Reference 5 with updates). Note the pronounced peak in the histogram around $a \sim 20,000$ AU. This peak was first noticed by Oort in 1950 (6) using a sample of only 19 cometary orbits. From such a histogram, Oort concluded that the solar system must be surrounded by a spherically symmetrical halo of comets, gravitationally bound to the Sun, and at distances greater than 10,000 AU. This halo of comets is currently called the Oort cloud. The position of its inner edge is quite uncertain, but the position of its outer edge is defined by the Sun's tidal truncation radius of 100,000–200,000 AU (7).

Oort realized that the typical change in energy that a comet receives when it passes through the planetary system (the part of the solar system within Pluto's orbit) is approximately $\pm 0.0005 \text{ AU}^{-1}$. Yet the peak in Fig. 2 is fairly narrow. It is unlikely that a comet that started out in the peak on its first passage into the planetary system would have a value of $1/a$ which would leave it in the peak on subsequent passages. Thus, we refer to comets where $a \geq 10,000$ AU as dynamically "new" comets because it is likely that they are on their first passage into the planetary system. Comets not in the peak in the histogram have most likely passed through the planetary system before. On a first passage into the inner solar system, gravitational perturbations from Jupiter would eject roughly half of the new comets and capture the other half to more tightly bound (smaller values of a), less eccentric orbits. Only 5% of new comets would be returned to Oort cloud distances (8).

Returning to the short-period comets, Fig. 3 plots the inclination i as a function of semimajor axis a for the 130 short-period comets that are currently cataloged. Note that the vast majority of these objects have semimajor axes between 3 and 4 AU. This group of objects tend to have their *aphelions* (farthest distance from the Sun) near Jupiter and have small encounter velocities with Jupiter. The orbits of this group of comets are highly influenced by Jupiter. These comets have come to be known as Jupiter-family comets (JFCs). Typically, there has been an arbitrary cutoff for JFCs at an orbital period of 20 years. This limit is

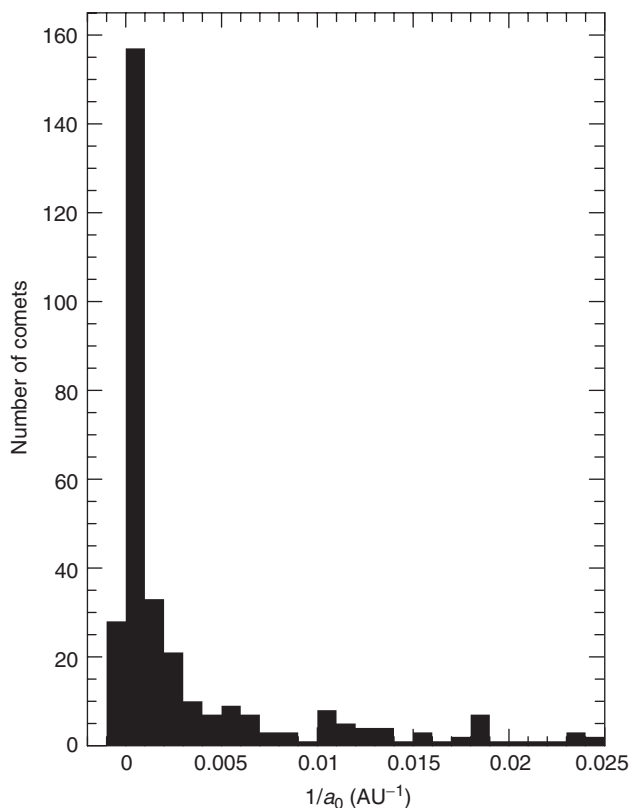


Figure 2. The number of comets as a function of inverse original semimajor axis. The data for this histogram are from Reference 5 with further updates from Marsden (personal communication). Note the sharp peak at $1/a_0$ slightly larger than 0. This is the evidence that there is a spherical reservoir of comets known as the Oort cloud.

shown as a dotted line in the plot. Comets whose orbital periods are between 20 and 200 years are referred to as Halley-type comets (HTCs), after their most famous member that has a 76-year orbital period.

In Fig. 3, note that all of the JFCs are in prograde orbits that have relatively low inclinations (a mean of $\sim 13^\circ$) and some of the HTCs are in retrograde orbits ($i > 90^\circ$; the HTCs have a mean inclination of $\sim 78^\circ$). These differences in properties suggest that the JFCs and HTCs may be dynamically different and perhaps have different origins.

Most of the JFCs are clustered at inclinations of less than 35° , suggesting that the three comets whose periods are less than 20 years but inclinations are greater than 40° may be different from the other JFCs. Carusi et al. (9) have offered an alternate, less arbitrary, definition for the division between JFCs and HTCs. They suggest that the boundary should be based on a parameter known as the Tisserand parameter T , which is defined as

$$T = a_J + 2\sqrt{(1 - e^2)a/a_J \cos(i)}, \quad (1)$$

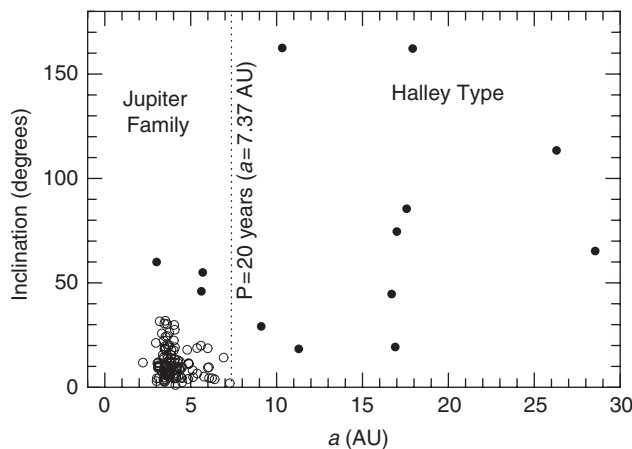


Figure 3. The inclination as a function of the semimajor axis for the short-period comets. In this figure, the orbits for the 130 cataloged short-period comets are shown. The dotted line divides the plot into two segments based on the traditional separation between Jupiter-family and Halley-type comets of orbital period = 20 years. The open circles are those comets whose Tisserand parameter (see text) $T > 2$; the closed circles are comets where $T < 2$. Jupiter-family comets are confined to the plane of the ecliptic, preferentially, whereas Halley-type comets are not.

where a_J is the semimajor axis of Jupiter's orbit and a , e , and i refer to the semimajor axis, eccentricity, and inclination of the cometary orbit. In their scheme, the division between JFCs and HTC's would be $T = 2$. The Tisserand parameter is a measure of the relative velocity between Jupiter and a comet during a close approach, here $v_{\text{rel}} \propto \sqrt{3 - T}$. Objects where $T > 3$ cannot cross Jupiter's orbit and are confined to orbits either totally interior to (the case for Comet Encke) or totally exterior to (the case for Chiron) Jupiter's orbit. In Fig. 3, the value of the Tisserand parameter is coded as an open circle for comets where $T > 2$, or JFCs, and closed circles for $T < 2$, or HTC's. Inspection of this figure now shows that the JFCs in this definition all have low inclinations with a mean of 11.6° , whereas the HTC's seem to have no preferred inclination. Note that this definition is an improvement over the 20-year period definition because it takes into account Jupiter's gravitational influence on the comets. Thus, this definition, along with some modifications (10), is now being widely adopted to differentiate between different types of comets.

Note that the orbits of the HTC's are reminiscent of the long-period and nonperiodic comets in that there are no preferred inclinations for their orbits. Thus, both of these types of comets appear to come from a spherical reservoir of a form different from the disk envisioned for forming the planets. JFCs, on the other hand, have low inclination orbits. For this reason, they are sometimes referred to as ecliptic comets.

How do comets get from the outer solar system to the inner solar system? The most important force acting on Oort cloud comets is the gravitational effects of the disk of our galaxy (11,12). The gravitational tidal forces of the galactic disk will cause the Oort cloud to be a prolate spheroid whose long axis points toward the center of the Galaxy. This force acts like a torque on each comet and leaves

the semimajor axis virtually unchanged but causes the perihelion distance (closest approach distance to the Sun) to undergo a random walk. When the perihelion distance crosses into the planetary system, the planets (especially Jupiter) begin to have a major influence on the orbit.

Other external forces that influence cometary orbits are passing giant molecular clouds (GMCs) or stars passing through the Oort cloud. These are much less important than the Galactic gravitational field. Passing GMCs are rare events that have a mean interval of occurrence of 300 million years (8). However, when they do occur, they have a major effect on the part of the Oort cloud they pass.

During the lifetime of the solar system, about 5500 stars have passed within 100,000 AU of the Sun. A star of the same mass as our Sun tunneling through the Oort cloud with a velocity of 20 km sec^{-1} will eject all comets within about 450 AU of the star. Passing stars have probably ejected about 10% of the total population of the Oort cloud (8).

It is estimated that the mean dynamic lifetime of comets in the Oort cloud is only about 60% of the age of the solar system. Thus, the Oort cloud must somehow be replenished for Oort cloud comets to exist still. One possibility for replenishment is the capture of comets that originally formed around other stars and then were ejected into interstellar space. This process is very unlikely because the relative velocity of these comets with respect to our solar system makes them difficult to capture.

A more likely scenario for replenishing the Oort cloud is that an inner Oort cloud exists at a few thousand AU from the Sun whose cometary orbits are perturbed to move them from the inner Oort cloud to the outer Oort cloud (13). When these inner Oort cloud comets are at aphelion, Galactic and stellar perturbations would be sufficient to raise their perihelia from the inner to the outer Oort cloud. However, normal effects of the Galactic gravitational field would not affect them other than at aphelion, so that the orbits of inner Oort cloud comets would otherwise be stable for the entire lifetime of the solar system. There is no physical demarcation for the inner and outer Oort clouds, and the differences between these clouds is only a matter of definition. The inner Oort cloud is, in general, just the part of the Oort cloud that remains relatively inactive, whereas the outer Oort cloud is the principal source of comets that enter the inner solar system.

Let us return to the question of why the Oort cloud appears spherical although we theorize that comets are leftover debris from the formation of the planets in a disk of material surrounding the young Sun. Current dynamic simulations have shown that the comets did not form originally at the location of the present Oort cloud. Instead, the comets are leftovers from the region of the giant planets, Jupiter through Neptune (14). Small objects that remained in the region of the giant planets after the planets formed would come under the gravitational influence of these planets. As a planetesimal comes near a planet, the planet exerts a gravitational tug on the planetesimal. This excess force is strongest at the closest approach to the larger body, but because the two bodies are in different orbits that have different velocities around the Sun, the excess force is exerted only for a short time. This excess force changes the orbit of the planetesimal (because of conservation of momentum, it also changes the orbit of the larger planet, but by a very much smaller amount than the change in the

planetesimal's orbit because the mass of the planet is much larger than the mass of the planetesimal). Thus, the orbit of the planetesimal is changed in a random-walk fashion. When the planet that causes the change in orbit is very massive, such as Jupiter or Saturn, then the excess velocity imparted to the planetesimal is most often large enough to eject the planetesimal from the solar system on a hyperbolic orbit. Some small fraction of objects would have their orbits changed to orbits near Uranus and Neptune or out to the inner Oort cloud.

Uranus and Neptune are not massive enough to impart such a large increase in velocity. Instead, the influence of Uranus and Neptune would increase eccentricities, perihelia, and inclinations of the orbits. Ultimately, most of the objects that start in the solar nebular disk at the distance of the current Uranus and Neptune would end up in orbits whose semimajor axes are several thousand AU but their orbits would no longer have inclinations that were small with respect to the plane of the ecliptic. Instead, the perturbations of the orbits by Uranus and Neptune would randomize the inclinations, and the planetesimals would form a spherical cloud at the distance of the inner Oort cloud. From there, they would populate the outer Oort cloud.

Thus, we are able to start with a disk of material in the protoplanetary system and, through the influence of gravity, the major planets perturb the orbits into the spherical Oort cloud. Most of the Oort cloud comets were formed in the Uranus/Neptune region, whereas Jupiter and Saturn ejected most of the leftover debris that existed near their orbits.

Let us return to the short-period comets. In Fig. 3, it was shown that there are two dynamic classes of short-period comets, the JFCs and HTC. Using the Tisserand parameter definition for delineating these two groups, we noted that the average inclination of the JFCs was low (11.6°) and no orbits had inclinations greater than 35° . This is quite reminiscent of the original protosolar disk. The HTCs, however, show no preferred inclinations. In 1988, Duncan et al. (15) showed, by computer simulations, that it was impossible to reproduce the low-inclination, short-period comet orbits from perturbations of the spherical Oort cloud because comets that enter the planetary system from the Oort cloud would preserve their random inclinations. Duncan et al. showed that a cometary source that has a low orbital inclination distribution was far more consistent with the observed orbits. They posited that a belt of comets must exist outside Neptune's orbit that were the source of the short-period comets. They named this reservoir the Kuiper belt because it was first suggested by Kuiper in 1951 and Edgeworth in 1949 that the disk of the protosolar nebula would not be truncated at Neptune's orbit. (Some scientists think that the reservoir for short-period comets should be named the Edgeworth–Kuiper belt, but this is not universally accepted.) Further computer simulations have shown that a significant fraction of the orbits of objects in the Kuiper belt would be stable for the lifetime of the solar system. Perturbations of the giant planets would be sufficient occasionally to stir up the objects near the inner edge of the Kuiper belt and bring these objects into the inner solar system. However, the inclinations of the perturbed objects would not be changed enough to lose the signature of a disk.

In summary, from dynamic considerations, our understanding of the structure of the outer solar system is that it consists of two parts. Starting from

around the orbit of Neptune, there is a disk of objects, known as the Kuiper belt. This disk has a thickness of order $\pm 25^\circ$. This disk is a primordial disk of objects (it formed at the same time as the planetary system), and a significant fraction of the orbits of objects in this belt is stable for the lifetime of the solar system. Starting at a few thousand AU from the Sun, there is a separate reservoir of comets known as the Oort cloud. This reservoir is spherical in shape, not disk-like. The Oort cloud consists of two regions, the inner and outer Oort clouds. The inner cloud is relatively unperturbed and replenishes the outer Oort cloud, but the orbits of objects in the outer Oort cloud are stable only for about 60% of the lifetime of the solar system. Objects in the Oort cloud did not form in the cloud but instead formed in the Uranus/Neptune zone, and planetary perturbations changed their orbits to the Oort cloud. Thus, the Oort cloud is not a primordial reservoir of comets.

Objects that reside in the Oort cloud are much too faint to be detected by current techniques. Our only evidence for the Oort cloud, therefore, is the peak in the histogram of the inverse semimajor axis of long-period comets shown in Fig. 2. Our concept of the mass of material in the Oort cloud comes about from simulations and is quite uncertain (estimates range from a few to 50 Earth masses).

We have hard observational evidence for the existence of the Kuiper belt. The first Kuiper belt object was discovered in 1992 (16) and since that time, 241 objects whose orbits are in the Kuiper belt have been discovered (as of 21 February 2000—more are being discovered regularly). The objects that have been discovered are relatively large; diameters range from 50 to 500 km. Even so, they are up to 100,000 times fainter than can be seen with the naked eye and require very difficult, dedicated searches. However, these objects are clearly much larger than the comets that are seen in the inner solar system as active comets. One study (17) using the Hubble Space Telescope to image 100 times fainter than ground-based surveys reported the discovery of objects of more typical cometary dimensions. However, this study is controversial and has not yet been duplicated.

Figure 4 shows the inclinations of the orbits of the current catalog of Kuiper belt objects as a function of the semimajor axis. Included on this plot are several dashed vertical lines that mark the location of orbits that resonate with Neptune's orbit. The 2:3 mean motion resonance refers to an orbit in which the object orbits the Sun twice for every three times that Neptune orbits the Sun (that is, the object is in general exterior to Neptune's orbit). These resonances are very special dynamic places. If an object starts out in orbit nearest to Neptune in a 2:3 resonance, then every time it orbits the Sun twice, it will come back nearest to Neptune. Thus, Neptune will have an excellent opportunity to perturb the object repeatedly, and this orbit is unlikely to remain stable for very long. On the other hand, if an object is in a 2:3 resonance and starts off away from Neptune, it can *never* get near Neptune, and it is protected from Neptune's influence, even if the eccentricity of the orbit is such that the object crosses Neptune's orbit. Pluto's orbit is an example of an orbit that crosses Neptune's orbit, but Pluto is in a 2:3 resonance with Neptune and thus, is protected. Inspection of Fig. 4 shows that many of the Kuiper belt objects that have been discovered are in protected resonances, especially those whose orbits cross Neptune's. Nonresonant orbits inside Neptune's orbit have been cleared of objects.

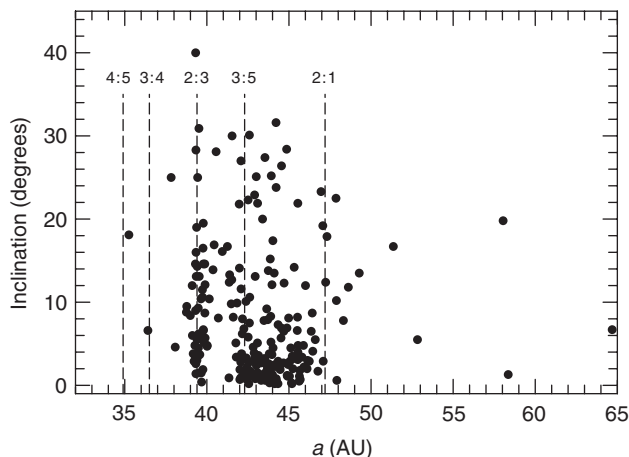


Figure 4. The inclination as a function of the semimajor axis for the known Kuiper belt objects. In this figure, 228 Kuiper belt orbits are shown. Thirteen objects have semimajor axes >55 AU. Mean motion resonances of Neptune are denoted by vertical dashed lines.

From the properties of the discovered objects, we can attempt to estimate the number and mass of Kuiper belt objects. While doing this, we must keep certain difficulties in mind. We can discover only the larger, brighter objects in the inner part of the Kuiper belt. Thus, we have no knowledge of the density of small objects, nor of the density of objects in the outer part of the Kuiper belt. Indeed, we have no real knowledge of the distance to the outer edge of the Kuiper belt. When objects are discovered, we can measure their brightness. From this, we need to figure out their sizes and masses. The Kuiper belt objects do not shine on their own, but, instead, they reflect light from the Sun. We know how much light the Sun emits at all wavelengths, and we know how far the objects are from the Sun. However, to convert from brightness to size, we need to know what percentage of the incoming light is reflected from the surface of the object. This percentage is known as the albedo. The albedo of these objects is generally unknown, though we have an idea of the albedo of some active comets, especially Halley's comet. Typical active comets have albedos of 4%, that is, only 4% of the incident light is reflected. If comets are ice, why is the albedo so low? After all, ice is bright white! Think in terms of snow that has fallen and been plowed and has had dirt mixed into it. That snow will be dirty and gray and much less bright than freshly fallen snow. The same is true of comets. The dirt in the nucleus mixes with the ice and makes the ice much less reflective than if it were pure. If we assume that comets in the Kuiper belt also have 4% albedo, we can convert from brightness to size. These estimates of size are only as good as our estimate of the albedo. Pluto has an albedo of 40%, which may be due to frost deposition, but may tell us something about albedos in the same region as the Kuiper belt. If Kuiper belt objects had similar albedos, they would be much smaller than our current estimates. From the size, we can determine the volume. Then, if we know the density of the body, we can determine the mass. However, as discussed above, the density of comets is very uncertain. Taking all of these factors into account, we believe that the Kuiper belt has only a fraction of an Earth mass of material.

Composition

To date, several spacecraft have visited comets and have made measurements *in situ*, but we have yet to retrieve samples and return them for study in our laboratories. The vast majority of our knowledge of cometary composition comes, however, from remote observations obtained with ground-based or Earth-orbital telescopes. These data are gathered using spectroscopic and photometric techniques on a wide variety of comets. The spectrum of a comet is characterized predominantly by molecular emission coupled with an underlying continuum of sunlight reflecting from the dust.

As indicated earlier, the nucleus of a comet is a combination of ices and refractory (dusty) material. About 50% of the mass of the comet is ice and the rest is dust. As the comet approaches the sun and sublimates and gas flows away from the nucleus into the coma, gas phase chemical reactions take place. Some of the chemical reactions are between two molecules or between molecules and dust, but the vast majority of the reactions are photochemical reactions resulting from the interaction of solar radiation and gases in the coma. The gas species become changed quickly. We term the original gaseous species that are the same as the ices the parent species. Species that are produced by chemical reactions of all types are termed daughter species (though some are granddaughter, great-granddaughter, etc.). Optical spectra of comets have been studied for more than 100 years. Until recently, only daughter species were detected. In the past two decades we have detected directly parent species with observations in the IR and mm. Table 1 contains a list of species that have been observed in the spectra of comets. Obviously, some of the species listed here are radicals and cannot be parent species. Others are stable species that could be parent species or could be the result of chemical reactions.

Gas. Spectra of comets can be used for determining the bulk composition of coma gases. However, it requires sophisticated models to use observations of daughter species to determine the nature of parent species. The importance of understanding the nature of ices is to understand which species were formed in the interstellar medium, before incorporation into the protosolar nebula and which species are the result of chemical processing in the early protosolar nebula. The exact composition of the species, the ratio of one species to another, and the isotopic ratios yield information about the conditions in which the ices formed.

Table 1. Atoms and Molecules Detected in Comets

Type species	Species
Atoms	H, O, C, S, Na, K, Ca, V, Cr, Mn, Fe, Co, Ni, Cu
Diatomics	OH, CH, NH, CN, C ₂ , CO, CS, S ₂ , SO
Polyatomics	NH ₂ , C ₃ , H ₂ O, CO ₂ , HCN, HNC, H ₂ S, OCS, H ₂ CO, H ₂ CS, CH ₄ , C ₂ H ₂ , C ₂ H ₆ , NH ₃ , HNCO, HCOOH, HC ₃ N, CH ₃ OH, CH ₃ CN, NH ₂ CHO, CHOOCH ₃
Ions	C ⁺ , Ca ⁺ , OH ⁺ , CH ⁺ , CO ⁺ , N ₂ ⁺ , H ₂ O ⁺ , HCO ⁺ , H ₃ O ⁺ , CO ₂ ⁺
Solids	Organics, silicates

Our understanding of the chemical composition of comets has increased tremendously in the past few decades because of improvements in our ability to observe comets with an ever increasing array of highly sensitive instruments. Originally, cometary spectra were confined to the optical region of the spectrum, using photographic plates as the detector. These observations were limited in spectral coverage and in detail because the dynamic range of plates is too limited to record properly the range of abundances in comets.

In the 1980s, we added charge-coupled devices (CCDs) to our arsenal of optical detectors. These allowed including the near-infrared region of the spectrum, and they possessed the dynamic range necessary to study coma densities that range across many orders of magnitude.

In the 1980s, the International Ultraviolet Explorer (IUE) spacecraft was launched. This spacecraft made it routine to observe comets in the ultraviolet, where several important atomic and molecular transitions, such as hydrogen Lyman α and Lyman β , CS, CO, S₂, etc., are found. Systematic studies of comets in the UV by the IUE yielded important understanding of the relative abundances in comets. These observations have been continued with the Hubble Space Telescope (HST) since the demise of the IUE. The HST is much more sensitive than the IUE and can cover the optical and IR wavelengths, in addition to the UV. It also has much higher spatial resolution and yields important information about the distribution of gases within the coma.

Radio observations of comets were added in the 1970s but lacked the sensitivity to observe all except the OH radical. In the 1980s, came more sensitive radio detectors and the advent of millimeter-wave observations of comets. In addition, sensitive IR array detectors were developed, allowing IR observations of comets. The radio, millimeter, submillimeter, and IR bandpasses are extremely critical to our understanding of comets because it is only in these bandpasses that we can detect parent molecules directly.

All of these detector improvements, coupled with the appearance of the bright comets Hyakutake and Hale–Bopp in the mid-1990s, have allowed a manyfold increase in our understanding of the chemistry of comets.

Inspection of Table 1 shows that a wide range of species is detected, from very simple atoms and molecules to quite complex molecules. However, 80% of the ices are water ice. The presence of fully oxygenated species alongside fully reduced species (i.e., CO versus CH₄ and C₂H₂) indicates that the comets could not form in a thermochemically equilibrated region of the solar nebula.

HNC is a species that could yield clues to the amount of chemical processing in the protosolar nebula (HCN is hydrogen cyanide, whereas HNC is hydrogen isocyanide, a differently bonded combination of the three atoms). HNC is seen in the interstellar medium at a ratio $\text{HNC}/\text{HCN} \propto 1/T$ (where T is the temperature of the interstellar region). HNC could be a parent molecule and would be indicative of the formation temperature. If HNC were a parent, then the ratio HNC/HCN would be constant as the heliocentric distance of the comet changed. Irvine et al. (18) measured both HNC and HCN in comet Hale–Bopp as a function of heliocentric distance and found that the ratio was not constant with heliocentric distance. They concluded that this was clear indication of ion–molecule chemistry in the coma and that HNC was not a parent and could not yield information on deposition temperatures.

Formaldehyde (H_2CO) is ubiquitous in the interstellar medium, and its presence in comets would be expected if comets incorporated interstellar ices. However, it is also possible to produce formaldehyde through chemical processing in the protosolar nebula, so the fact that we detect formaldehyde in comets does not necessarily indicate that ices survived from the interstellar medium until they were incorporated in comets. The detailed abundance of formaldehyde can lend a clue to this question. However, we have evidence that the formaldehyde may be contained in inhomogeneous inclusions in the nucleus, so our present understanding of the quantity of H_2CO is insufficient to determine its origins.

Earth's atmosphere is more than 70% nitrogen. This nitrogen is important for life. How does the nitrogen abundance of comets compare with that of Earth? This is mostly an unanswered question because nitrogen-bearing species are very difficult to measure. The most likely nitrogen-bearing parent species in a comet are the dust, which does have some nitrogen, N_2 , NH_3 , and HCN . Of these, only HCN is easily detected, but it is expected to be a trace species. Wyckoff et al. (19) reviewed the various observations that bear on the question of the abundance of nitrogen in comets and concluded that the total nitrogen relative to other species is depleted from solar by about 6–10 times. Until more data become available, this is still an open question.

The frosts of noble gases are extremely volatile, so observations of noble gases would represent very sensitive thermometers of the formation temperatures and subsequent evolution of comets. Measurements of the noble gases would indicate the warmest temperatures the comet had ever reached. Unfortunately, noble gases are extremely difficult to detect, and none was identified in the *in situ* measurements of Halley's comet. So far, only weak upper limits have been obtained for any noble gases (20).

Much can be learned about the fractionation history of the solar nebula by looking at the isotopic ratios of the comets. (Isotopes of an element have the same number of protons but different numbers of neutrons in their nuclei, e.g., hydrogen and deuterium). The fractionation history is a function of temperature, density, and pressure. If we can determine the isotopic ratios as a function of *formational* distance from the Sun, we have accurate values for these parameters. In general, the isotopic ratios for comets bear much more resemblance to those of the Sun and meteorites than they do to the values in the interstellar medium.

The isotopic ratio D/H is of particular importance to our understanding of the evolution of the solar nebula and its bodies, and particularly to our understanding of Earth. The protosolar nebula started with a value of D/H which was representative of the interstellar medium at the time it was formed. Various processes cause the loss of deuterium and hydrogen from a planet's atmosphere. Less massive bodies lose more of their light elements as they evolve because their gravitational fields are smaller than the more massive bodies. Because hydrogen is lighter than deuterium, hydrogen is lost first. It is, thus, easy to deplete both of these elements but almost impossible to enrich them. Because H is lost faster than D, the ratio of D/H will tend to increase as a result of loss processes in a planet. Material at different heliocentric distances will also have different D/H values because of the temperature gradients in the solar nebula. Figure 5 shows measured values for D/H in various solar system bodies. These are difficult

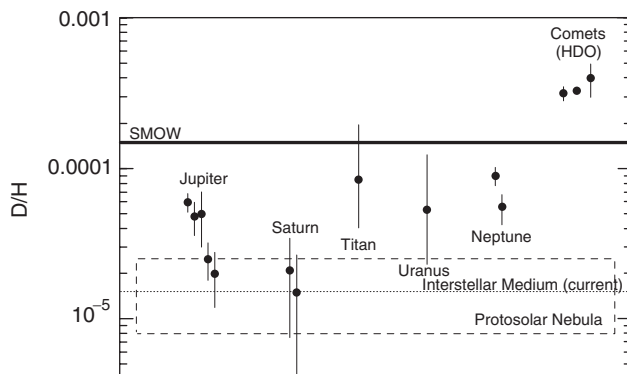


Figure 5. The D/H ratio of a number of solar system bodies. The data for the giant planets and Titan come from Table 3 of Lecluse et al. (34). The comet values are, from left to right, for Halley (35), Hyakutake (36) and Hale–Bopp (37). Also marked on the plot are values for D/H in the interstellar medium and a range of values for D/H from protosolar nebula models. A heavy line marks the D/H ratio for standard mean ocean water (SMOW), a measure from Earth’s oceans. For details of the individual numbers, see the sources cited.

measurements, so for some bodies there are multiple values, and they do not always agree. As one moves upward on this plot, one is measuring an enriched D/H value compared with the original protosolar value.

In Fig. 5, D/H values are plotted for three different comets. The agreement in their ratios is quite good, but a great deal more data are necessary before we can understand the degree of homogeneity in the comet formation region. Recall that comets formed in two distinctly different regions of the protosolar nebula. All three comets are Oort cloud comets, and the agreement suggests that the value of D/H measured may be typical of Oort cloud comets. It will be necessary to sample comets from both reservoirs to understand the conditions better. However, no Kuiper belt comets are bright enough to measure the D/H ratio with current equipment. As detectors and telescopes improve, this situation should change. Note, too, that the plotted values for cometary D/H are obtained from observing H_2O and HDO. For comet Hale–Bopp, the D/H ratio was also measured by observing HCN and DCN. The ratio obtained this way was about a factor of 7 higher than that measured using the water species. We will return to this later.

The D/H ratio of Earth is measured from oceanic water and is shown on the plot as a heavy line marked SMOW (standard mean ocean water). Note that it is heavily enriched compared to the protosolar nebula. Comets show a value about twice the value for SMOW. It has been suggested that the enrichment over protosolar measured for Earth results from comets having been the source of the water on Earth. Although it is sure that a number of comets have bombarded the earth, thus contributing to the water content of Earth, the fact that the D/H ratio of comets is twice as high as Earth’s means that comets could not have been the most significant source of water on Earth. If comets were a major contributor to the oceans, the D/H ratio for SMOW would be closer to the D/H ratio for comets. Thus, D/H ratios can point to comets as only a minor source of Earth’s oceans and limits the rate of bombardment by cometary bodies.

In addition to measurements of D/H in comets, isotopic ratios of $^{12}\text{C}/^{13}\text{C}$ (in the gas) for about half a dozen comets and $^{14}\text{N}/^{15}\text{N}$ (for Hale-Bopp) have been measured. These isotopic ratios are similar to the value measured in the Sun and are vastly different from interstellar values. Measurements of $^{12}\text{C}/^{13}\text{C}$ in dust particles collected by the various Halley flybys show that this ratio is highly variable in the dust.

We have gathered various pieces of evidence; some of them indicate that comets contain material from the interstellar medium, but others indicate that there was substantial chemical processing in the protosolar nebula. Can any of these pieces of evidence indicate the temperature at which the cometary ices were deposited? Several lines of evidence yield such deposition temperatures.

The ratio of $\text{H}_2\text{O}/\text{CO}$ is indicative of the deposition temperature. However, extreme caution must be taken when computing this ratio because of the relative volatility of the two species. If observations are made of a comet at 5AU, then any sublimation is being driven by the CO, not the H_2O . However, at 1AU, both species are sublimated, so a measure of $\text{H}_2\text{O}/\text{CO}$ gas at 1 AU will yield a value indicative of the ratio in the ices and thus gives information on sublimation temperatures. Laboratory experiments have been conducted by Bar-Nun and co-workers to determine the order in which the ices are deposited on grains and the subsequent ratio of sublimed gas. Comparison with observations of Halley, Hale-Bopp, and Hyakutake indicate that all three of these comets have deposition temperatures of ~ 50 K.

Recall that the D/H ratio, as measured by DCN, was a factor of 7 higher than the D/H ratio measured for HDO. The disparity between these ratios is also an indicator of the deposition temperature and implies a deposition temperature of 25–50 K. This is consistent with the $\text{H}_2\text{O}/\text{CO}$ temperature given before.

Dust. Before the spacecraft flybys of comet Halley in 1986, our picture of the dust component in comets was entirely shaped by indirect evidence. The color of the continuum and the structure of the dust tail gave indications of the sizes of the particles in the coma. In addition, it was considered likely that some dust particles collected in Earth's upper stratosphere originated from comets, although it was impossible to link any of the collected particles to a particular comet.

Our picture of the dust was that it was a component entirely separate from the gas. Our ideas changed dramatically with comet Halley and with subsequent IR observations of other comets. Three of the five spacecraft that visited comet Halley had instruments on them to measure the dust. Giotto, Vega 1, and Vega 2 all carried instruments to measure the composition of the cometary dust by having the dust hit a target surface and performing time-of-flight spectroscopy on the resultant components. These instruments were called PIA (particle impact analyzer) on Giotto and PUMA on the two Vega spacecraft. Based on these experiments, it was determined that the dust component of the coma had two different parts: an organic component rich in carbon, hydrogen, oxygen, and nitrogen (dubbed CHON particles) made up around 30% of the sample; Mg-rich silicate particles constituted another third of the particles; the remainder of the particles consisted of a mix of the two types of particles (21). The CHON particles, it was shown, are fragile and showed evidence of fragmentation during their outflow into the coma. The volatile nature of these grains allows them to be

converted into gaseous species as they flow outward. Thus, the CHON particles turned out to be an extended source of gas in the coma.

The PIA/PUMA instruments sampled only small grains in the range of 10^{-16} – 10^{-11} g. This would correspond to spherical grains 0.06–3 μm in diameter. However, grains this size represent only a small fraction of the total dust mass. Another instrument on the Giotto spacecraft, DIDSY (dust impact detection system), measured grains as large as 10^{-5} kg (22). In addition, during the close approach of the Giotto spacecraft to the comet, an impact of a single grain on the spacecraft sent the spacecraft wobbling and damaged the optics. Models of the necessary size of the impactor indicate that the grains were as large as $\sim 2 \times 10^{-2}$ kg.

The PIA/PUMA instruments measured the mass of the detected particles, their densities, and their compositions. They found that the CHON particles had a density of $\sim 1 \text{ g cm}^{-3}$ and the silicate particles had a density of $\sim 2.5 \text{ g cm}^{-3}$. This implies that the CHON-dominated grains are quite fluffy, whereas the silicate-dominated grains are more compact. However, the heavier grains of a particular type are as fluffy as the lighter grains. The bulk composition of the dust analyzed by PIA/PUMA shows that the dust in Halley's coma has an elemental composition which is similar to the solar system composition (of course, the dust is depleted in hydrogen from the solar value but that is because it is easily lost, as explained before). This implies that Halley's comet was not fractionated with respect to the protosolar nebula (21).

The analysis of the composition of individual grains shows a lack of chemical equilibrium among the grains. This was deduced from examining the ratios $\text{Fe}/(\text{Fe} + \text{Mg})$ and Fe/Si versus Mg/Si , two quantities that have been used for laboratory samples of dust to distinguish different types of particles. The analysis also found very few, if any, Ca-Al-rich grains in Halley's dust. One startling discovery was that the isotopic ratio $^{12}\text{C}/^{13}\text{C}$ was not the same at the grain level but varied from 1 to 5000 for individual grains (the normal solar system value is 89, and meteorites have ranges from 4–120). The low cometary dust values may be an instrumental artifact, but the high values appear to be real. The only explanation for such high $^{12}\text{C}/^{13}\text{C}$ ratios is that these particular grains were produced in a nucleosynthetic process, probably before the protosolar nebula was formed, and were incorporated into the comet unchanged.

Recently, the Infrared Space Observatory (ISO) observed the IR spectrum of comet Hale-Bopp in the spectral region from 7–45 μm . The spectrum obtained when the comet was at around 3 AU shows features that can be matched well by crystalline magnesium-rich olivine (fosterite, Mg_2SiO_4) (23). Other ISO spectra, obtained when the comet was closer to the Sun, might also contain features of pyroxene. In the interstellar medium, any observed silicates are only in their amorphous form, so the detection of *crystalline* fosterite is quite significant. At the temperatures and pressures of comet formation, silicates would form in their amorphous state, not in a crystalline form, so the detection of crystalline silicates in comets requires that these crystalline silicates formed elsewhere and were incorporated into comets when already in their crystalline form. ISO also detected, for the first time, crystalline silicates in evolved stars (so-called AGB stars and planetary nebulae), as well as in young stellar objects (24). This suggests that the crystalline silicates might have been formed before

the protosolar nebula was created and were incorporated into comets when they were formed.

Diversity. Detailed study of individual comets such as Halley or Hale-Bopp tells us much about one comet. However, to characterize the early protosolar nebula, we need to study a large sample of comets and determine if the compositions are similar. If all comets have the same composition, this indicates that the protosolar nebula in the region where the comets formed was homogeneous. If all of the comets are not the same, then there could be a primordial difference, implying that the protosolar nebula was lumpy, or there could be an evolutionary difference. Extensive ground-based telescopic surveys have been conducted to determine the degree of heterogeneity of cometary abundances. These studies were carried out by Cochran and co-workers at The University of Texas, McDonald Observatory; by Newburn and Spinrad at The University of California, Lick Observatory; and the largest survey was completed by A'Hearn, Millis and co-workers, primarily at Lowell Observatory.

A'Hearn et al. (25) compared observations of 85 comets and looked at relative gas abundances. They found that the vast majority of comets seems to have similar relative abundance (one species relative to another). However, they found that a group of comets exists that is depleted in carbon-chain molecules, such as C_2 and C_3 , but not in other carbon-bearing molecules, such as CN (comet Giacobini-Zinner is the prototype for this group of depleted comets). Although the carbon-chain-depleted comets represent a minority of their sample, all of the carbon-chain-depleted comets are Jupiter-family comets. Recall that JFCs formed *in situ* in the Kuiper belt, but other comets formed interior to the Kuiper belt and were ejected to the Oort cloud. Thus, if all of the carbon-chain-depleted comets are JFCs and none of the other comets is depleted, this points to differences in the conditions in the two zones of formation. It should be noted that not all JFCs show the carbon-chain-molecule depletion. This still leaves the question whether there are Oort cloud comet interlopers in the JFC sample, whether we are seeing compositional differences in the Kuiper belt region, or whether there are evolutionary differences in Kuiper belt comets.

Evolution of Cometary Nuclei

Comets represent the least altered components left over from the formation of the solar system, but they are probably not totally pristine. Though the Oort cloud is sparsely populated, collisions between Oort cloud comets can still occur occasionally. Several physical processes can also modify the outer layers of cometary nuclei while they reside in the Oort cloud.

The ices of the nucleus are subjected to various forms of radiation, including the radiative flux from passing stars, UV flux from forming stars while the protosolar nebula was still in a star-forming region, X-ray photons from the Galactic background and nearby exploding supernovae, and Galactic cosmic rays. These forms of radiation can profoundly alter the nature of the ices by altering the chemical bonds, radiation darkening and polymerizing ices (26,27). The radiation damage can increase the density and porosity of the ices and can turn simple ices into more complex ones. These processes can penetrate to

different depths within the nucleus; UV-induced changes occur in only the outermost layers (a few to hundreds of atoms or 1–10 nm deep), and X-ray induced changes as deep as 10–100 μm . The depth of the alteration for charged particle radiation would depend on the energy of the particles; low-energy solar protons hardly penetrate, and high-energy protons penetrate deeply. If cometary surfaces have low density, the radiation damaged crust might extend many meters deep (28). Depending on the depth of radiation penetration, the altered ices may survive one or more passages into the inner solar system before being lost to sublimation and subsequent coma outflow.

Collisions between comets are responsible both for increasing the size of objects through accretion and reducing some objects to dust via grinding. Calculations by Stern (29,30) show the frequency of collisions in each region. The Kuiper belt comets which have already been discovered are large and have volumes some 1000–10,000 times that of a Halley-sized comet. To build each object, it was necessary for many collisions to take place (30). Comet–comet collisions in the Oort cloud are less important than in the Kuiper belt and probably only 1 in 10,000 comets suffers a catastrophic collision in the age of the solar system (29). However, small debris in the region (some of which is the result of previous collisions) will collide with the comets, gardening, polluting, and roughening the surface.

When the comets pass into the inner solar system and are heated, sublimation of the gas will carry away much of the embedded dust. However, some of the dust grains are sufficiently large that they cannot escape even the weak gravitational field of the nucleus. These grains will fall back upon the surface. In time, unless some event catastrophically ejects the entire surface, these larger grains will form a mantle on the surface of the comet, insulating the ices and choking the ability of the ices to sublime (31). Indeed, this scenario of sealing the surface of the nucleus has been proposed as one ending for the life of the comet.

Cometary Impacts

Every single day, about 100 tons of interplanetary material hits the upper atmosphere of Earth, and some of that material reaches the surface. A large percentage of the smallest particles originates in comets as dust grains that are released when ices sublime. These small cometary particles represent no hazard to Earth. Many of them are fluffy aggregates that “float” in the upper stratosphere for long periods of time and are occasionally collected for study by specially equipped high-flying aircraft.

The present gentle rain of small particles on the upper atmosphere has not always been the pattern for impacts on Earth. In the early stages of the formation of the solar system, many planetesimals still remained, and the young solar system was a much more chaotic place. Before they were ejected from the solar system, the icy planetesimals’ orbits were perturbed so that they often impacted the planets in the inner solar system. We see this record on the surfaces of atmosphereless and geologically inactive bodies such as the Moon or Mercury. Soon after the planets were first formed, they were being bombarded by icy and

rocky planetesimals (comet and asteroid pieces) at a rate which was about 1000 times greater than the current rate; the bombardment rate fell dramatically to “only” 30 times the current rate around 3.5 billion years ago (32). While the planets were being so heavily bombarded, they were in constant upheaval. It was impossible for life to form. Around 3.8 billion years ago, a period we call the late heavy bombardment, conditions began to change for the better for life. The oldest currently known fossils that indicate life forms are about 3.5 billion years old. Thus, the heavy rate of cometary impacts must have frustrated the origins of life in the earliest days of the solar system.

Did comets also serve as a catalyst for life on Earth? Comets contain large quantities of water, a necessary ingredient for life, along with other more volatile materials. It has been proposed that these organics, along with water, may represent the important prebiotic building blocks necessary for life (33). The arrival of large quantities of cometary material in the early evolution of Earth may have made available necessary raw materials for the formation of amino acids. Thus, once the late heavy bombardment had ended, the materials which were delivered to Earth from the bombardment might have played a critical role in the origins of life on Earth.

Once life had begun, the future interactions of Earth and comets was not entirely beneficial or even benign. We know that, even in recent times, objects impact Earth, as evidenced by locales such as the Barringer meteor crater in Arizona, the Chicxulub crater in the Yucatan (the remains of the impact event which, it is thought, destroyed the dinosaurs at the time of the Cretaceous–Tertiary (K/T) boundary 65 million years ago), or the Tunguska impact zone where an impactor fell in 1908. Many of these impactors are *not* icy, cometary bodies, but are asteroidal in origin. Rocky materials are stronger than the fragile cometary nuclei and can more easily survive passage through the atmosphere to impact the ground. However, comet impacts are still possible. A recent spectacular example of a cometary impact was the Shoemaker–Levy 9 impacts (in 1994) on Jupiter. Though these impacts had no effect on Earth, they serve as a reminder of the magnitude of effects that are caused by impacts on planetary bodies. Indeed, cometary impacts can be more extreme relative to the mass of the impactor than asteroidal impacts because of the random orientations of the cometary orbits relative to the plane of the ecliptic. If an impact occurs with the impactor in a retrograde orbit, the relative velocity of the comet and the Earth can be very large, and the impact energy is proportional to the square of that relative velocity.

The best defense against impactors of any type is to know in advance that the body exists and is potentially hazardous. NASA is currently coordinating large search efforts for near-Earth objects to catalog as many objects as possible. This effort is likely to be relatively complete for bodies of asteroidal origins, but many cometary bodies will go undetected in these searches because the random inclinations of the cometary orbits make systematic searches difficult and expensive. Thus, we could be faced at any time with the threat of a cometary impact and the resultant damage. Our best hopes for planetary protection are to be as vigilant as possible and to enhance our understanding of the nature of cometary nuclei so that we understand the properties of potential impactors.

Missions To Comets

Much can be learned about the nature of comets via observations from the ground and Earth orbit. However, we can never see the nucleus of the comet in this way, and we must infer information about its nature by using models. To study the nuclei of comets, it is necessary to send spacecraft to them. These spacecraft missions can range from simple flybys, to rendezvous missions, to landings and sample returns. In this section, the completed and planned cometary missions of the U.S. and other nations are described.

Completed Missions

ICE. The first cometary mission did not start its life as a mission to a comet. The International Sun Earth Explorer (ISEE) 3 was a mission to study the interaction of the solar wind with Earth's magnetosphere. It was launched in 1978. After excursions into Earth's magnetotail and five swingbys of the Moon, the spacecraft was retargeted to flyby comet Giacobini-Zinner and its name was changed to the International Cometary Explorer (ICE). The flyby of comet Giacobini-Zinner occurred on 11 September 1985, making this mission the first to encounter a comet.

Because the mission was intended to study the plasma environment of Earth, no cameras were carried by the spacecraft. The goal of ICE was to study the interaction of a comet with the solar wind. The spacecraft was targeted to fly through the tails of Giacobini-Zinner to gather data on the changing magnetic field and plasma. The experiments on board ICE found a thin, dense, and cold plasma sheet. The magnetic field strength was low, and the spacecraft detected a change in polarity as it crossed the comet axis. The temperature of the plasma increased with increasing distance down the tail until about 20,000 km from the nucleus. Because the mission was not intended to sample a comet, very little compositional information was obtained.

Sakigake, Susei, Vega 1, Vega 2, and Giotto. Comet Halley is one of the most famous comets, and so it was not surprising that there was great interest in missions to this comet. An international armada of spacecraft visited this comet: the European Space Agency sent the Giotto mission; the USSR sent two spacecraft, Vega 1 and 2; and Japan sent two spacecraft, Sakigake and Susei. A detailed description of these missions and their first science results can be found in a special issue of *Nature* (Volume 321, May 1986).

The two Japanese spacecraft were officially part of the Planet-A project of the Institute of Space and Astronautical Science (ISAS); one spacecraft was an engineering test spacecraft dubbed MS-T5 (renamed Sakigake, Japanese for "forerunner", after launch), and the other the Planet-A main spacecraft (renamed Susei, Japanese for "comet", after launch). They were launched in January and August 1985, respectively. The two spacecraft were identical but carried different instrument complements: Sakigake carried a plasma-wave probe, whereas Susei carried an ultraviolet imager and charged particle analyzer. Concerns about the dust environment near the nucleus caused the probes to be targeted well sunward of the nucleus; Sakigake passed 7 million km away on 11 March 1986, and Susei 151,000 km away on 8 March 1986. The relative velocity of Susei and the comet was 73 km sec^{-1} .

The two Vega missions were combination missions, whose goals were studying Venus with balloons and landers on their way to a rendezvous with comet

Halley. They were launched 6 days apart in December 1984 and flew by Venus 4 days apart in June 1985. They flew by comet Halley on 6 March and 9 March 1986 at around 8,000 km on the sunward side of the nucleus. The two three-axis stabilized spacecraft were identical, including instrumentation. Each carried a complement of 15 scientific instruments to measure the plasma, the dust, the composition, and to obtain images of the comet. The encounter velocities were 79.2 and 76.8 km sec⁻¹.

The Giotto mission was launched by the European Space Agency (ESA) on 2 July 1985; it was the first ESA interplanetary mission. It achieved closest approach to the nucleus on 14 March 1986 at an encounter distance of 600 km and a flyby velocity of 68.4 km sec⁻¹. The spacecraft was spin-stabilized and carried a complement of 10 instruments, including a camera, mass spectrometers, dust detectors, and plasma instruments. The close encounter distance coupled with the fast flyby velocity were a calculated risk to the spacecraft but were driven by the desire to get close for the best imaging and for sampling by the various instruments. The biggest risk came if the spacecraft entered a dust jet (a dust jet is a region of the coma in which the flux of 1- to 10- μ m particles is enhanced by a factor of 3–10 over the ambient coma). Fourteen seconds before the closest approach, the spacecraft apparently was hit by a “large” dust grain, causing the spacecraft’s angular momentum vector to shift by 0.9° and to wobble. Because the spacecraft had no onboard data storage devices, this interrupted communications and data transfer, though some data were transmitted intermittently for the next 32 min. Some of the instruments were damaged at this time, but in general, the spacecraft remained healthy. Thus, ESA was able to retarget the spacecraft for a flyby of comet Grigg–Skjellerup on 10 July 1992, but some instruments, such as the cameras, were not functional.

The Giotto mission’s Halley multicolor camera achieved the milestone of the first excellent images of the nucleus of a comet, supplemented by images from the Vega cameras. The images showed a lumpy, elongated object; the dimensions were 8 × 8 × 16 km with several very active jets, and the rest of the surface was relatively quiescent; only 20–25% of the surface appears to be active. The nucleus had a very dark albedo of 2–4%, comparable to the darkest known bodies in the solar system. The comet does not rotate simply around a single axis but has a complex wobble-spin tumbling motion. The dust environment was quite severe, especially at the high encounter velocities. It was this armada of spacecraft that finally confirmed our general picture that 80% of cometary ice is H₂O ice.

Deep Space 1. This mission is the first of the NASA New Millennium missions intended primarily as demonstrators of cutting-edge technology never before flown. The Deep Space 1 mission was launched on 24 October 1998. It was primarily an engineering mission that had a science component planned to be executed if possible. This mission was intended to test 12 new technologies; the most important was ion propulsion, but included such technologies as autonomous navigation, low power electronics and new instruments. The nominal scientific mission was a flyby of asteroid 1992 KD (Braille) at 10-km altitude on 28 July 1999. The mission was extended to flyby comet 19P/Borrelly on 22 September 2001 at a closest distance of 2170 km from the nucleus. During the course of the flyby, 52 visible-wavelength images and 45 infrared spectra were obtained.

Deep Space 1 observed an irregularly-shaped cometary body with a long axis of 8 km and a 2:1 axis ratio (38). The surface had a low albedo of 3% and true albedo variations from 1–3%. Only about 10% of the surface appeared to show activity and no ices were detected on the surface. A jet appeared to emanate directly from the rotational pole.

Current Missions. Great interest continues in understanding comets and their constraints for models of the protosolar nebula. Thus, several missions are currently planned and launched or to be launched within the next decade. This section will describe all of these missions.

Stardust. This is the fourth mission in the NASA Discovery program of small focused planetary science missions, and it is the first robotic return of material from beyond the Moon's orbit. The Stardust spacecraft was launched on 7 February 1999 (Principal Investigator Dr. Donald Brownlee of the University of Washington). The Stardust mission will encounter comet Wild 2 in June 2004 at a relative encounter velocity of 6.1 km sec^{-1} and an encounter distance as close as 150 km. Wild 2 is expected to be much less active than comet Halley, and the much lower encounter velocity should protect the spacecraft. In addition, the encounter with Wild 2 will take place at a heliocentric distance of 1.86 AU, where comets are less active than the 0.9 AU heliocentric distance of Halley at encounter. Stardust will carry a camera for imaging and a dust particle analyzer (CIDA—a similar instrument to the PIA/PUMA instruments discussed earlier).

The most important goal for Stardust is to capture samples of interplanetary dust (collected in 2000) and cometary dust and return them to Earth for analysis. It will accomplish this goal by using a capture target built of *aerogel* attached to panels on the spacecraft. Aerogel is an ultra-low-density microporous substance which has been tested on previous shuttle missions. Once the flyby of the comet has been achieved, the aerogel target will be closed into a sealed sample return capsule (SRC), and the spacecraft will return to the vicinity of Earth. When the spacecraft approaches Earth, the SRC will be released from the spacecraft and will descend into the atmosphere on a parachute. It is planned that the sample will land in the Utah Test and Training Range, where the SRC will be recovered by helicopter or ground vehicle and transported to the planetary materials curatorial facility at the Johnson Space Center. The Earth return of the SRC is scheduled for 2006. More information can be found at <http://stardust.jpl.nasa.gov>.

CONTOUR. The comet nucleus tour (CONTOUR) is another NASA Discovery mission; its goal is studying the diversity of cometary nuclei (Principal Investigator Dr. Joseph Veverka of Cornell University). CONTOUR includes detailed investigation of two diverse short-period comets: Encke (November 2003) and Schwassmann–Wachmann 3 (June 2006). However, the mission profile is extremely flexible and can be modified to include the first-ever study of a “new” comet (such as Hale–Bopp) should one be discovered during the mission.

The spacecraft is scheduled for launch in July 2002 and will carry four instruments with it: a wide-field and a high-resolution camera, a dust analyzer (CIDA—a copy of the Stardust instrument), and a neutral and ion mass spectrometer. To maintain its extreme orbital flexibility, the CONTOUR spacecraft uses repeated Earth-return trajectories and multiple gravity-assist maneuvers. A unique feature of CONTOUR is that it will be placed into a spin-stabilized

“hibernation” mode, requiring no ground contact, during its cruise intervals between comet encounters and Earth swingby maneuvers. The spacecraft will be “awake” for 75 days around encounters and 50 days around each Earth maneuver. More information about CONTOUR can be found at <http://www.contour2002.org/>.

Rosetta. The Rosetta mission is the third *Cornerstone* mission of the European Space Agency (ESA). This mission is to be launched January 2003 from Kourou, French Guiana, aboard an Ariane-5 rocket. Rosetta is designed to rendezvous with comet P/Wirtanen in August 2011, orbit around the comet, making observations of the nucleus, and eventually, to land a package of instruments called the Rosetta Lander (or RoLand) on its surface in August 2012. The long cruise phase (eight years) is necessary to rendezvous with the comet far from the Sun and at a low enough encounter velocity.

Rosetta is larger than the previous three missions. It carries 12 instruments in addition to RoLand. The orbiter instruments include a camera, several spectrographs, a neutral and ion mass spectrometer, a dust analyzer similar to that on Stardust and CONTOUR, a dust detector, and a plasma instrument. RoLand carries a suite of eight additional instruments to evaluate samples of the comet *in situ*. No samples will be returned to Earth. RoLand will include an instrument to take a core sample from 200 mm under the nucleus’ surface.

One unique problem faced by the Rosetta mission is to land successfully on the surface when we do not currently have any information about the nature of the material on which RoLand will land. RoLand will land with a relative velocity of less than 1 m sec^{-1} and will fire an anchoring harpoon into the nucleus to secure the lander to the surface (recall that the gravity of a comet is weak because it has such low mass). Once tethered to the surface, RoLand will begin its measurements, transmitting the data to the orbiting spacecraft for relay to Earth. More information can be obtained about Rosetta at <http://sci.esa.int/rosetta/>.

Deep Impact. The final cometary mission is another in the Discovery program. Deep Impact is to be launched in January 2004 and will encounter comet Tempel 1 on 4 July 2005 (Principal Investigator Dr. Michael A’Hearn of the University of Maryland). Deep Impact’s goal is to sample the deep interior of a cometary nucleus. It would be quite impossible to do this by trying to drill a core sample, so the Deep Impact approach is quite different. Deep Impact will expose material deep in the interior by creating a crater of $> 100\text{-m}$ diameter and $> 20\text{-m}$ deep. This will be done by releasing an autonomously targeted mass of roughly $1/2$ metric ton (mostly composed of copper) one day before the impact. Cratering is a natural process in space. The difference between a naturally created crater and that created by the Deep Impact spacecraft is that we will know the size and mass of the impactor and exactly when it hit.

The impactor will be carried by the flyby spacecraft. After the release of the impactor, the flyby spacecraft will decelerate slightly so that its closest approach to the nucleus will occur after the impact at an interval roughly twice the predicted time to form the crater. The impactor will carry a camera to obtain high-resolution images of the surface of the impact site just before impact. The flyby spacecraft will carry two cameras and two spectrometers to observe all phases of the crater-forming process, the resultant crater, the debris plume, and the

changes induced in outgassing. The comet will be observed simultaneously from Earth to record data about the changing comet. For more information, see <http://www.ss.astro.umd.edu/deepimpact/>.

Summary

Comets are small icy/dusty bodies left over from the formation of the solar system and represent the least altered bodies from the epoch of formation. They formed in the outer parts of our planetary system; some still reside in their primordial reservoir known as the Kuiper belt; those formed in the Uranus-Neptune region were ejected to a spherical reservoir known as the Oort cloud; those formed near Jupiter and Saturn were ejected from our solar system entirely.

The nuclei of comets are small (typically a few km) bodies composed of ices and dust. Of the ices 80% is water ice. The dust is composed of organics and silicates. As the nucleus is heated, the ices sublime and flow outward from the nucleus. The dust/gas mixture forms a large coma and even larger tails.

We are on the verge of increasing our knowledge of comets tremendously with a series of spacecraft already launched and about to be launched. The next decade should revolutionize our understanding of comets and thus, the solar nebula.

BIBLIOGRAPHY

1. Shu, F., J. Najita, D. Galli, E. Ostriker, and S. Lizano. The collapse of clouds and the formation and evolution of stars and disks. In E.H. Levy, and J.I. Lunine (eds) *Protostars and Planets III*. The University of Arizona Press, Tucson, AZ, 1993, p. 3.
2. Beckwith, S.V.W., and A.I. Sargent. Circumstellar disks and the search for neighboring planetary systems. *Nature* 383: 139 (1996).
3. Whipple, F.L. A comet model: The acceleration of comet Encke. *Astrophys. J.* 111: 375 (1950).
4. Hainaut, O., R.M. West, B.G. Marsden, A. Smette, and K. Meech. Post-perihelion observations of comet P/Halley. IV. $R = 16.6$ and 18.8 AU. *Astron. Astrophys.* 293: 941 (1995).
5. Marsden, B.G. *Catalogue of Cometary Orbits*, 12th ed. Smithsonian Astrophysical Observatory, Cambridge, MA, 1997.
6. Oort, J.H. The structure of the cometary cloud surrounding the solar system and a hypothesis concerning its origin. *Bull. Astron. Neth.* 11: 91 (1950).
7. Smoluchowski, R., and M. Torbett. The boundary of the solar system. *Nature* 311: 38 (1984).
8. Weissman, P.R. Dynamical history of the Oort cloud. In R.L. Newburn, Jr., M. Neugebauer, and J. Rahe (eds), *Comets in the Post-Halley Era*. Kluwer Academic, Dordrecht, The Netherlands, 1991, p. 463.
9. Carusi, A., L. Kresak, E. Perozzi, and G.B. Valsecchi. High-order librations of Halley-type comets. *Astron Astrophys.* 187: 899 (1987).
10. Levison, H.F. Comet taxonomy. In T.W. Rettig, and J.M. Hahn (eds), *Completing the Inventory of the Solar System*. Ast. Soc. Pacific. Conf. Series 107, San Francisco, CA, 1996, p. 173.
11. Heisler, J., and S. Tremaine. The influence of the Galactic tidal field on the Oort comet cloud. *Icarus* 65: 13 (1986).

12. Bailey, M.E. The mean energy transfer rate to comets in the Oort cloud and implications for cometary origins. *Mon. Not. R. Astron. Soc.* 218: 1 (1986).
13. Hills, J.G. Comet showers and the steady-state infall of comets from the Oort cloud. *Astron. J.* 86: 1730 (1981).
14. Duncan, M., T. Quinn, and S. Tremaine. The formation and extent of the solar system comet cloud. *Astron. J.* 94: 1330 (1987).
15. Duncan, M., T. Quinn, and S. Tremaine. The origin of short-period comets. *Astrophys. J. (Letters)* 328: L69 (1988).
16. Jewitt, D., and J. Luu. Discovery of the candidate Kuiper belt object 1992 QB₁. *Nature* 362: 730 (1993).
17. Cochran, A.L., H.F. Levison, S.A. Stern, and M.J. Duncan. The discovery of Halley-sized Kuiper belt objects using the Hubble Space Telescope. *Astrophys. J.* 455: 342 (1995).
18. Irvine, W.M., J.E. Dickens, A.J. Lovell, F.P. Schloerb, M. Senay, E.A. Bergin, D. Jewitt, and H.E. Matthews. Chemistry in cometary comae. *Faraday Discuss.* 109: 475 (1998).
19. Wyckoff, S., S.C. Tegler, and L. Engel. Nitrogen abundance in comet Halley. *Astrophys. J.* 367: 641 (1991).
20. Stern, S.A., J.C. Green, C. Cash, and T.A. Cook. Helium and argon abundance constraints and the thermal evolution of comet Austin (1989c1). *Icarus* 95: 157 (1992).
21. Jessberger, E.K., and J. Kissel. Chemical properties of cometary dust a note on carbon isotopes. In R.L. Newburn, Jr., M. Neugebauer, and J. Rahe (eds), *Comets in the Post-Halley Era*. Kluwer Academic P, Dordrecht, The Netherlands, 1991, p. 1075.
22. McDonnell, J.A.M., P.L. Lamy, and G.S. Pankiewicz. Physical properties of cometary dust. In R.L. Newburn, Jr., M. Neugebauer, and J. Rahe (eds), *Comets in the Post-Halley Era*. Kluwer Academic, Dordrecht, The Netherlands, 1991, p. 1043.
23. Crovisier, J., K. Leech, D. Bockelee-Morvan, T.Y. Brooke, M.S. Hanner, B. Altieri, H.U. Keller, and E. Lellouch. The spectrum of comet Hale-Bopp (C/1995 01) observed with the Infrared Space Observatory at 2.9 AU from the Sun. *Science* 275: 1904 (1997).
24. Waelkens, C., L.B.F.M. Waters, M.S. De Graauw, E. Huygen, K. Malfait, H. Plets, B. Vandenbussche, D.A. Beintema, D.R. Boxhoorn, H.J. Habing, A.M. Heras, D.J.M. Kester, F. Lahuis, P.W. Morris, P.R. Roelfsema, A. Salama, R. Siebenmorgen, N.R. Trams, N.R. Van Der Blik, E.A. Valentijn, and P.R. Wesselius. SWS observations of young main-sequence stars with dusty circumstellar disks. *Astrophys. J.* 315: L245 (1996).
25. A'Hearn, M.F., R.L. Millis, D.G. Schleicher, D.J. Osip, and P.V. Birch. The ensemble properties of comets: Results from narrowband photometry of 85 comets, 1976–1992. *Icarus* 118: 223 (1995).
26. Andronico, G., G.A. Baratta, F. Spinella, and G. Strazzula. Optical evolution of laboratory-produced organics: Applications to Phoebe, Iapetus, outer belt asteroids and cometary nuclei. *Astro. Astrophys.* 184: 333 (1987).
27. Strazzula, G., and R.E. Johnson. Irradiation effects on comets and cometary debris. In R.L. Newburn, Jr., M. Neugebauer, and J. Rahe (eds), *Comets in the Post-Halley Era*. Kluwer Academic, Dordrecht, The Netherlands, 1991, p. 243.
28. Mumma, M.J., P.R. Weissman, and S.A. Stern. Comets and the origin of the solar system: Reading the Rosetta stone. In E. Levy, and J. Lunine (eds), *Protostars and Planets III*. The University of Arizona Press, Tucson, AZ, 1993, p. 1172.
29. Stern, S.A. Collisions in the Oort cloud. *Icarus* 73: 499 (1988).
30. Stern, S.A. Collisional time scales in the Kuiper disk and their implications. *Astron. J.* 110: 856 (1995).
31. Brin, G.D., and D.A. Mendis. Dust release and mantle development in comets. *Astrophys. J.* 229: 402 (1979).

32. Hartmann, W.K. Paleocratering of the Moon: Review of post-Apollo data. *Astron. Space Sci.* 17: 48 (1972).
33. Chyba, C.F., and C. Sagan. Comets as a source of prebiotic organic molecules for the early Earth. In P.J. Thomas, C.F. Chyba, and C.P. McKay (eds), *Comets and the Origin and Evolution of Life*. Springer, New York, 1997, p. 147.
34. Lecluse, C., F. Robert, D. Gautier, and M. Guiraud. Deuterium enrichment in giant planets. *Planetary Space Sci.* 44: 1579 (1996).
35. Eberhardt, P., M. Reber, D. Krankowsky, and R.R. Hodges. The D/H and $^{18}\text{O}/^{16}\text{O}$ ratios in water from comet P/Halley. *Astron. Astrophys.* 302: 301 (1995).
36. Bockelee-Morvan, D., D. Gautier, D.C. Lis, K. Young, J. Keene, T. Phillips, T. Owen, J. Crovisier, P.F. Goldsmith, E.A. Bergin, D. Despois, and A. Wooten. Deuterated water in comet C/1996 B2 (Hyakutake) and its implications for the origin of comets. *Icarus*. 133: 147 (1998).
37. Meier, R., T. Owen, H.E. Matthews, D.C. Jewitt, D. Bockelee-Morvan, N. Biver, J. Crovisier, and D. Gautier. A determination of the HDO/H₂O ratio in comet C/1995 O1 (Hale-Bopp). *Science*. 279: 842 (1998).
38. Soderblom, L.A., T.L. Becker, G. Bennett, D.C. Boice, D.T. Britt, R.H. Brown, B.J. Buratti, C. Isbell, B. Glese, T. Hare, M.D. Hicks, E. Howington-Kraus, R.L. Kirk, M. Lee, R.M. Nelson, J. Oberst, T.C. Owen, M.D. Rayman, B.R. Sandel, S.A. Stern, N. Thomas, and R.V. Yelle. Observations of comet 19p/Borrelly by the Miniature Integratee Camera and Spectrometer aboard Deep Space 1. *Science*. 296: 1087 (2002).

ANITA L. COCHRAN

The University of Texas McDonald Observatory
Austin, Texas

COMMERCIAL APPLICATIONS OF COMMUNICATIONS SATELLITE TECHNOLOGY

Introduction

According to the Teal Group, satellite communications industry consultants, some 1,017 commercial satellites will be launched in the 10 years from 1999 to 2008. Of these, 102 satellites will be dedicated to conventional broadcasting of radio, television, and data: 82 to direct-to-home satellite broadcasting; 384 to broadband and multimedia transmissions; and 449 to mobile communications. Together, these satellites represent a total market value of US \$49.8 billion (1).

Commercial communication satellites constitute the fastest growing segment of the global satellite market, including satellites for the military civilian government, and scientific applications. This rapid growth is driven by the development of new and more cost-efficient satellite technologies and the demand for more telecommunications services worldwide.

Transmission of network, independent, and cable television programming to affiliates and cable headends will continue to be a dominant application of

commercial satellites. The U.S. company PanAmSat, based in Greenwich, Connecticut, for example, has 22 satellites orbiting Earth that perform this service, as well as satellite newsgathering (SNG) and transmitting live special events from around the world. The company, owns the world's largest commercial geostationary satellite system and is the world's largest provider of video and data broadcasting services via satellite. PanAmSat plans to expand its global fleet to at least 25 spacecraft by mid-2001. On 1 January, 2000—and on the days immediately preceding and following this date—PanAmSat delivered millenium celebrations to viewers throughout the world for more than 40 customers, including ABC, BBC, CNN, Fox, and NBC. In total, there were more than 1300 hours of live event coverage and some 300 satellite feeds. PanAmSat used a total of 12 U.S. domestic and international satellites plus 5 teleport facilities.

Since 1980, with the debut of international broadcaster Cable News Network (CNN), satellite newsgathering has been an increasingly large user of satellite time because of the world's growing appetite for live TV coverage of breaking news. During the past 20 years, satellite technology has become a primary tool of newsgathering for television news organizations throughout the world because fiber connectivity is often difficult or impossible to arrange on short notice and from remote locations. The excitement of real-time news from anywhere is something that television viewers throughout the world have come to expect, either through local stations or news organizations like CNN. CNN is the world's first 24-hour-a-day global news channel and is now seen in 184 million households in more than 210 countries. In recent years, CNN and other TV reporting from such venues as the Persian Gulf, Bosnia, and Chechnya have enables viewers worldwide to share live experiences and to form their own opinions based on what they hear and see via satellite. Competition among broadcasters continues to stimulate SNG. For example, the American television network ABC has about 80 SNG trucks deployed around the world.

Most SNG transmissions use the higher frequency Ku band rather than the C band, which is used for most broadcast and cable program distributions to system affiliates. There are several reasons for this. First, unlike C-band transmissions, Ku-band transmissions are largely unaffected by terrestrial microwave interference and large structures in metropolitan areas. Second, the cost of Ku-band satellite dishes for SNG broadcasting is only one-fourth of that for C-band dishes. And third, Ku-band dishes are at least three times smaller than traditional C-band dishes, making them easier to transport and set up quickly.

Another factor stimulating the growth of SNG is the technology of digital compression, one of the manipulations of data made possible by digitalization. Through digital compression, it is possible to use only a fraction of a satellite transponder for transmission. Thus, the lower costs associated with SNG have helped to expand the number of SNG broadcasts around the world.

In fact, digital technology does much more than merely allow for compression of television signals, although this is a major commercial capability. Digital television allows for CD-quality audio and cinematic-quality video, both of which generate files so dense with data that they cannot be transmitted across the airwaves without digital video and audio compression. Digital satellite technology is also making possible the transmission of high-definition television because an HDTV picture requires much more information than a conventional analog

television channel can provide. In the coming decade, as consumers buy new digital television sets to replace their aging analog models and as TV production houses acquire digital cameras and other digital technologies for their studios, many more programs will be produced in the HDTV format.

As full implementation of digital technology and the compression of TV channels occur, virtually each new satellite will be able to broadcast up to six times as many programs simultaneously, using the same amount of bandwidth, compared with the older analog technology. This is putting satellites on the same footing as cable in terms of their capability to deliver hundreds of channels into the home. In fact, satellite-to-home television companies like EchoStar Communications and DIRECTV are challenging cable's dominance and keeping cable subscriber costs down, where satellite services can easily compete. Furthermore, a measure passed by Congress in 1999 allowing satellite TV operators to provide local broadcast TV channels to their subscribers is expected to produce huge market gains for the satellite TV industry (2). It is projected that, by 2003, more North American TV households will receive digital TV signals via satellite than by cable (3). DIRECTV now delivers more than 225 digital satellite TV channels directly into the home.

Direct broadcast satellites (DBS) are a worldwide phenomenon. Today, throughout the world, more than 62 million ground receivers measuring as little as 18 inches (45 centimeters) in diameter look at more than 60 high-powered geostationary satellites that deliver direct-to-home (DTH) programming. Of these satellite dish antennas, 42.7 million are in 36 countries of Europe, North Africa, and the Middle East (4). The phenomenon of DBS was first solidly established in Europe in the early 1990s by such providers as British Sky Broadcasting (UK), SES Astra (Luxembourg), and Telenor (Norway). By the mid-1990s, the skies above Asia were also populated with direct-to-home broadcast satellites. For example, since 1990, AsiaSat of Hong Kong has launched two satellites that cover 66% of the world's population in 53 countries that span Asia, India, and the Middle East. In the United States, four companies pioneered DBS: EchoStar, Primestar, United States Satellite Broadcasting, and DIRECTV. Subsequently, DIRECTV absorbed both Primestar and USSB, and now is the largest US DTH television company. In the coming years, DBS growth overseas is expected to exceed U.S. growth; Europe will contribute 40% of all DBS subscribers by year-end 2008 (5). In the shorter term, between 2000 and 2003, the number of DBS subscribers will triple (6).

Interactive TV is another powerful commercial application made available by digital television technology. Viewers at home can participate in live programs and manipulate the broadcasts: they can choose camera angles, stop, reverse, or go back to viewing a program in real time. Companies such as WebTV, TiVo, Sun Microsystems, Intel, and America Online (AOL) are forging new relationships with broadcasters and satellite operators that will showcase their new interactive services. For example, at the Consumer Electronics Show in Las Vegas in January 2000, TiVo, which has partnered with Hughes Electronics Company's DIRECTV service, demonstrated TiVo's VCR-like software that can pause, rewind, and fast-forward live and recorded television.

Also, Hughes and AOL formed a strategic partnership to link the satellite operator with a leading content provider. In 1999, AOL invested US \$1.5 billion

in Hughes to get its interactive multimedia AOL-TV product onto the DIRECTV platform and launch its AOL-Plus high-speed Internet service via DirecPC, a high data rate service from Hughes. AOL-TV will allow users to surf the Web, "chat" online, and E-mail back and forth as they watch TV. Finally, a deal struck in June 2000 between Microsoft and DIRECTV gives Microsoft's interactive WebTV product a place on the DIRECTV platform. The two companies will offer a new VCR-like service that marries a computer hard drive and a digital TV so that viewers can record up to 30 hours of TV programming for later playback. The service, called Ultimate TV, is similar to that offered by TiVo. In addition to digital video recording, it also offers Internet access and interactive TV.

Meanwhile, Sun Microsystems has demonstrated an interactive music program from the BBC that enables viewers to vote for their favorite bands and change camera angles while watching a weekly pop music show. And Intel and WebTV possess enhanced TV technologies that can overlay statistics onto sports broadcasts, add detailed product information to commercials, and let viewers play along with game shows. Commercial interactive and enhanced TV services will generate nearly US \$1 billion in revenues and reach more than five million households by 2002 (7). According to Forrester Research, within 5 years, interactive TV-based Web browsing will generate US \$11 billion in advertising and billions more in commerce (8).

The digitalization of the television signal, along with the fast growth of DBS, has encourage the worldwide radio industry to launch new, satellite-delivered domestic, regional, and international radio networks that provide CD-quality sound. Two of the new services Sirius and XM Radio, are scheduled to debut hundreds of channels of enhanced audio entertainment and information programming to fixed and mobile consumers from year-end 2000 to early 2001. A third service, WorldSpace, is already operational. Its goal is to provide high-quality, multilingual digital audio programming to the emerging and underserved markets of the world. The first WorldSpace satellite to become operational (in June 2000), AsiaStar, broadcasts more than 50 channels of programming to 4-inch satellite dish antennas. Two additional commercial satellites for Africa and Latin and South America will complete the WorldStar offering.

Data transmission is big business. A team of World Trade Organization (WTO) economists has estimated that, by year-end 2000, there will be more than 300 million Internet users worldwide and that electronic commerce will amount to a US \$ 300 billion a year industry. Beyond this, 80% of all business-to-business transactions will be conducted through the Internet by 2002 (9).

High-speed, interactive data transmission, which will allow real-time exchange of digital video and multimedia, is projected to become more widely available in the year-end 2001–2002 time frame. In the interim, companies such as Hughes Network Systems, Scientific-Atlanta, and Gilat Satellite Networks offer to business enterprises of all sizes satellite-based intra- and intercorporate networks that may extend across the city or around the world. Called VSAT (very small aperture terminal) networks—in recognition of their easily installed 24-inch satellite dishes—they can be used for a variety of purposes, including videoconferencing, employee and customer training, credit card authorization at the point of sale, new product introductions to marketing employees and

retailers, inventory control, and electronic data interchange (EDI) for ordering merchandise from manufacturers and paying for it. There are more than 500,000 VSATs installed in the world, including those in underdeveloped nations where they may be used to provide remote villagers their first access to telephony.

For example, in a remote village in Thailand, a community telephone booth with a small VSAT dish on top can transmit and receive phone calls to and from any person or place on Earth. Operated by the Telephone Organization of Thailand (TOT), this telephone is one of 4000 remote terminals plus 21 public switched telephone network (PSTN) gateways located in Thailand's provinces. This network is the world's largest VSAT network dedicated to telephony. It is an example of the way satellites can provide telephony service where the existing wire-line telecommunications infrastructure is limited or non-existent and/or where the population density is too low and the terrain too challenging to build a commercially viable wire-line system for rural telephony.

Once inside the PSTN, telephone calls can be routed anywhere in the world by wire-line, underwater cable, satellite, and/or fiber to the intended recipient. By using multiple gateways, the network can route calls to the nearest entry point in the PSTN, thus minimizing connection times and consumer costs. The Thai network is used mainly to provide coin-operated pay phones and private lines in villages and small communities around the country. Like other such VSAT networks in Southeast Asia and Africa and throughout the developing world, satellite telephony via VSAT networks is connecting small businesses with the global marketplace, giving them the opportunity to transmit two-way data as well as voice, no matter what the distance or terrain. In the coming years, most telephone VSAT networks will be replaced by new "bandwidth-on-demand" broadband satellite systems (see later), whose wide and spot beam antennas will provide ubiquitous coverage of Earth.

Because of the vast and ever-increasing worldwide use of VSAT and Internet applications, it is sometimes said that the initials "www" stand not for the World Wide Web, but for the World Wide Wait. At this time, there is simply not enough bandwidth to meet user demands—from cable, from telephone lines, or from satellites. However, this challenge should be resolved in the next year or two, when a number of new satellite broadband companies are scheduled to begin operations. This is none too soon in a situation where world-wide demand for digital bandwidth continues to surge. Consider China, where the number of Internet users, now 10 million (June 2000), is doubling every 6 months. By some calculations, China will have the second largest population of Web users in the world, after the United States, by 2005 (10). Overall, in Asia there are currently (June 2000) 40 million Internet users. This number is expected to grow to 375 million by 2005. Meanwhile, in the United States, the market for satellite-based Internet services has grown by 314% during the last year and by 858% during the last 2 years (11).

Despite laudable efforts by cable and telephony organizations worldwide to convert their narrowband telecommunications infrastructures to broadband speed (defined as at least 1.544 megabits per second), satellite companies are accomplishing this more quickly and at lower costs. For example, U.S. telephone service provider SBC Communications is in the midst of a 3-year program to build a US \$6 billion fiber broadband system that will cover only the western

United States. By comparison, current estimates indicate that a satellite broadband system covering the entire United States will cost US \$1.4 billion (12).

At least seven satellite companies, Astrolink LLC, @ Home in the Sky, EuroSkyWay, INMARSAT's Broadband Global Area Network (B-GAN), iSky, Spaceway, and Teledesic, plan to provide broadband multimedia services from new Ka-band satellites to customers in the Americas, Europe, Africa, Asia, and the Middle East. Another company, called SkyBridge, will provide broadband services using the Ku band, a set of frequencies that is much used today, principally by commercial TV broadcasters. SkyBridge has engineered its system so that no interference will be created by frequency sharing.

It is projected that business-to-business electronic exchanges alone will constitute a US \$1.1 trillion market worldwide by 2003 (13). Also stimulating this growth will be the new, consumer-oriented concept of "bandwidth on demand," which enables users to pay only for the amount of satellite bandwidth they actually use. Their link into the satellite system will be via a low-cost satellite dish as small as 26 in. (66 cm) in diameter. It is likely that, together, bandwidth on demand plus low access charges will draw cost-conscious consumers and businesses away from cable and telephone connections for their broadband requirements. There is an additional feature of a bandwidth-on-demand satellite system. Because capacity is released only when needed, satellite systems will be able to support a much higher number of users compared with cable and wire-line technologies.

The launch of these next-generation Ka- and Ku-band broadband systems is expected to usher in a "golden age" for the satellite industry (14). The new systems could enhance broadcasting with data and multimedia services, deliver Internet content, and deliver business communications with high-power spot beams that can provide two-way, real-time communications. It is projected that revenues from broadband-over-satellite services will rise from US \$25 million in 2000 to US \$21 billion in 2007 (15).

As VSAT networks for rural telephony continue to be built throughout the world, new regional and global handheld systems for mobile satellite telephony are now being demonstrated. Proponents of mobile handheld satellite telephony have stated that data delivery by phone will be the next big commercial boom among consumers. Already, individuals can be in touch by voice virtually anywhere in the world anytime. In addition, their handheld phone sets can also enable E-mail, Internet access, and one-on-one videoconferencing with a small, built-in TV screen.

According to the financial services company, Morgan Stanley Dean Witter, satellite-based mobile phone use could explode from a few thousand users today to around 17 million by 2007 (16). It is projected that, nongeostationary (GEO) systems will produce revenues of US \$62 billion in the next 5 to 10 years (17). How quickly that money will materialize, or if it will materialize at all, is a question open to much debate after the failure of Iridium, a US \$5 billion, low Earth orbit satellite global telephony system of 66 satellites that ceased operations in February 2000, not long after its start-up in November 1998. Industry analysts attribute Iridium's fate to large and heavy phone sets that did not operate indoors or in cars and heavy capital expenditures that necessitated high usage fees of up to US \$7 per minute.

Iridium did, however, successfully demonstrate its pioneering technology. The 66 satellites were cross-linked and could perform onboard switching to relay signals directly to and from handheld phones where that was feasible. The satellites acted like cellular towers in the sky, wireless signals moved overhead instead of through ground-based cells. The Iridium network also integrated land-based phone lines and local cellular facilities to provide the quickest and most efficient call routing.

As of May 2000, a new global mobile satellite telephony system started operations, and another project will start up early in 2001. Both the Globalstar and New ICO systems (described in detail later) will not be as technologically sophisticated (or costly) as that of Iridium. Both will make more use of existing terrestrial systems, so that the telephony signals will not be beamed up to or down from the satellite directly. Each will use existing terrestrial wire-line and cellular phone paths for originating and receiving calls.

In the meantime, four regional—as opposed to global—mobile satellite telephony systems are either planning to serve or are already serving customers in North America, Asia, the Middle East, Europe, and Africa. Dual-mode cellular/satellite mobile phones provide access to a vast array of consumer services and serve as the basic connection for the new “office in the car.” For example, the 2000 Ferrari model 550 Maranello is a prototype equipped with all of the latest in multimedia and navigation gadgets. The car’s cockpit is equipped with a personal computer hooked up to a swiveling satellite antenna to navigate the Internet; send and receive E-mail and exchange documents; download road maps; read weather forecasts and traffic news; book hotels, theaters, and restaurants; and even obtain bank balances and the latest stock quotes. This prototype was created to show the potential of EuroSkyWay, the first European broadband satellite network completely reserved for interactive multimedia services. Developed by Alenia Aerospazio, the system is scheduled to begin operations in 2002. EuroSkyWay’s satellites will allow real-time video communications for videoconferencing, high-quality image transfer, and access to the Internet that is 30 times faster than terrestrial modes, according to the company. EuroSkyWay’s target customers are business people who need to be in touch with the office and cannot stop working, even while in traffic (18).

In the coming years, satellite antennas mounted on car roofs will become virtually omnipresent because of important and potentially life-saving new technologies first introduced by the U.S. Department of Defense and now in full development by the commercial sector. For example, General Motors (GM) now equips many of its car models with a dual cellular/satellite communications system called OnStar™. Using OnStar, drivers can navigate efficiently to their destinations with a series of maps and audio commands. More importantly, wherever they are, they can contact local police, fire, and rescue departments, as well as AAA operators, and their precise geographical location will be relayed simultaneously via satellite. This is accomplished by using the Global Positioning System (GPS), a constellation of 24 satellites that has evolved well beyond its U.S. military origins. For the U.S. military, GPS provides situational awareness for troops in the field anywhere in the world and precision weapon guidance. Through GPS-based data, voice, and video communications, military personnel can interact with commanders at headquarters and receive real-time maps of

allied and enemy positions, weather maps and forecasts, and progress images of guided weaponry advancing toward their targets.

Besides GM, Ford, DaimlerChrysler, and other car manufacturers also provide “telematics” systems, “telematics” is the buzzword for the convergence of wireless technology, global satellite positioning, and onboard electronics in automobiles. But to date, the General Motors system is the most technologically advanced. For example, if keys are locked inside the car, OnStar will unlock the doors via satellite. OnStar also offers a voice-activated Internet access package that, via an onboard computer, enables sending and receiving data, voice, imaging, and E-mail. OnStar is expected to have four million users by 2002, and there is analyst speculation that GM will spin off OnStar as a separate tracking stock.

As of May 2000, there were four million GPS users worldwide. The market for GPS applications is expected to double by 2003, rising from US \$8 billion in revenues to US \$16 billion. This can be attributed in part to the fast-growing demand for GPS features to be outfitted in cell phones (19). As of June 2000, the GPS feature will be incorporated in as many as 18 satellites that are awaiting launch or are in production. The U.S. government currently provides this service free of charge to users worldwide.

As shown, the worldwide demand for bandwidth is growing exponentially. Among the chief factors fueling the need for bandwidth is the skyrocketing demand for two-way, high-speed Internet access. Despite the large investments made by cable and telephone industries to expand their network capacities, the demand for high-speed bandwidth far outstrips the supply (20). Satellites will be used extensively to take up the slack and to provide unique services that only satellites can provide because of their ubiquitous reach and cost-insensitivity to distance.

As can be seen, new commercial satellite communications applications are developing rapidly, fueled by new technologies. Most of these applications will require major capital investments and involve significant technical and business risks. It is likely that some of the new applications discussed in this article—written in mid-2000—will undergo major changes or will even fail. Iridium is an example. It is equally likely that new applications not discussed here will develop. It is certain that commercial satellite communications applications will continue to evolve rapidly. The following sections provide a road map for these developments.

Digital Satellite Television Broadcasting

As mentioned previously (see article in this volume on Communications Satellites, Technology of), television programs in the United States have been transmitted from network hubs to affiliates and cable headends since the 1970s. In recent years, a growing number of U.S. households also receive TV broadcasts directly from satellites, using rooftop antennas for Ku-band satellites and backyard antennas for C-band transmission. The United States is not alone in using satellites to broadcast TV programming. A 1999 annual Eutelsat survey of 36 countries in Europe, North Africa, and large parts of the Middle East showed

that 107 million households also get satellite-delivered TV, either through a dish or a local cable company. This is up from 95 million homes in 1998. A total of 42.7 million households in these regions are dish homes (21). Asian direct-to-home broadcasts also are plentiful, due in great part to Rupert Murdoch's Star TV satellite system. In fact, News Corporation (of which Rupert Murdoch is chairman) controls 100% of Star TV, 40% of British Sky Broadcasting, and 36% of Sky Latin America. All are DTH systems, and there are more than 100 million subscribers on these platforms alone. The number of dish homes around the world will continue to grow quickly because of the 800 million TV households in the world, nearly 700 million are in regions that have limited access to multichannel television. Terrestrial systems, such as cable or multipoint microwave distribution service (MMDS), are too costly and time-consuming to install, compared with direct satellite transmission.

Today, around the world, more than 62 million ground receivers that are as small as 18 inches (45 centimeters) in diameter look at more than 60 high-powered geostationary satellites that deliver DTH TV. Many of the DTH satellites can deliver more than 100 channels of premium digital video entertainment and information directly to consumer households. For example, since 1990, AsiaSat of Hong Kong launched two satellites that cover 66% of the world's population in 53 countries that span Asia, India, and the Middle East. AsiaSat's biggest customer for its satellites is programmer STAR TV, also of Hong Kong. Along with Asian-originated programming, STAR and other Asian DTH broadcasters use AsiaSat's satellites (along with those of competitor APT Satellite Company, Ltd.) to transmit such American TV fare as HBO, MTV, ESPN, CNN, and The Disney Channel, along with the programming of the BBC and other major international programmers.

DBS is thriving throughout the developed world. In Japan, Sky PerfecTV!, which uses two high-power JCSAT satellites for its DBS service, has more than two million subscribers. Malaysia is served by a fully digital DTH TV service that is carried on two MEASAT (Malaysia-East Asia Satellite System) satellites. The coverage pattern also includes the region from Malaysia to the Philippines and from Beijing to Indonesia. In November 1998, Russia's first commercial communication satellite, BONUM-1, was launched from Cape Canaveral, Florida, and is providing multiple-language DBS TV in European Russia, the Urals, and western Siberia. In Central and South America and the Caribbean, the Hughes Galaxy Latin America DIRECTV service delivers up to 232 channels of programming to more than one million subscribers in 27 countries.

Two EUTELSAT satellites launched in 2000 will meet market demands for new digital TV platforms and Internet Protocol (IP)-based services in Russia and Africa. Nineteen transponders on EUTELSAT W4 constitute a high-power fixed beam across Russia. Sixteen of these transponders are used for DTH digital TV broadcasting by the Russian media group Media Most. Twelve transponders are pointed over sub-Saharan Africa, where they are used for digital pay TV and broadband Internet access (22). EUTELSAT is Europe's leading satellite operator and ranks as one of the largest globally. It reaches across Europe, large parts of Africa and the Middle East and has connectivity with North America. Currently, EUTELSAT satellites can broadcast more than 600 analog and digital channels to more than 81 million satellite and cable homes. In addition to DBS,

EUTELSAT satellites are used for high-speed Internet connections, Internet backbone traffic, inter- and intracorporate networks, SNG, telephony, and mobile voice, data, and positioning services.

In the United States, two companies dominate the DTH satellite services market: DIRECTV from Hughes Electronics Corporation in EI Segundo, California, and EchoStar Communications Corporation in Littleton, Colorado. At the end of 1999, DIRECTV, which debuted in 1994, offered some 210 channels of programming from four proprietary DBS satellites. The company estimates it will have 10 million subscribers by 2001. In addition, DIRECTV also offers a service called DIRECTV Para Todos (For Everyone), a Spanish-language programming service that consists of 30 additional channels. DIRECTV Para Todos is aimed at the growing number of bilingual households in the United States. Through alliances with America Online, TiVo, and Wink Communications, DIRECTV plans shortly to introduce a portfolio of new interactive and data-enhanced television services, including data-enhanced video, electronic commerce, webcasting, software downloads, and two-way Internet access. DIRECTV is also helping to accelerate the advent of high-definition TV by introducing this year two channels of HDTV programming and integrated DIRECTV/HDTV television sets and set-top electronics.

EchoStar, which began operations in 1990 with large C-band satellite backyard antennas, today broadcasts from five proprietary DBS satellites under the trademark DISH Network. These five satellites give DISH Network the capacity for more than 500 channels of digital video delivered to homes in the continental United States via a single satellite dish. As of year-end 1999, DISH Network boasted more than three million subscribers. At the same time, EchoStar introduced DISHPlayer, which it claims is the world's first combined interactive Internet device, satellite TV receiver, personal video recorder, and game player. DISHPlayer was developed jointly with Microsoft's WebTV Networks.

As described before, by the early 2000s, we will see the second generation of digital direct-to-home TV, which will have a satellite return path and Internet access. As mentioned earlier, through this new interactive system, consumers will be able to access companion data for sports, documentaries, news, and other programming. For example, viewers watching CNN could use their remote control devices to request the latest sports scores and local weather forecasts. Viewers of live sports telecasts will be able to access up-to-the-minute player and team statistics. When an advertiser's commercial's broadcast, viewers will be able to use their remotes to order more information or a coupon for the product or the product itself. Also in the early 2000s, digital satellite TV receivers will be built into newly manufactured TV sets, making these TV sets "satellite ready." They will require only an antenna to receive multichannel, even multinational, TV.

The 1999 Satellite Home Viewer Act is a major factor that will increase DTH broadcasting growth in the United States. This U.S. Federal Communications Commission (FCC) regulation, signed into law by President Clinton, mandates that local broadcasters make their local and network signals available to DBS providers. The intent of the regulation is to increase competition in the cable industry and to keep cable TV rates "reasonable," according to the FCC. In 1999, satellite TV reached about 14 million U.S. households, compared with about 70 million homes reached by cable service. It is projected that by 2003 more

North American TV households will receive digital TV signals via satellite than by cable. Moreover, it is projected that the number of DBS subscribers will triple between 2000 and 2003 (23).

Two satellite technologies, in particular, will fuel the growth of DBS: (1) spot beams aboard Ku-band satellites, and (2) Ka-band satellites. Current satellite technology will allow designing a spot beam satellite that has up to 50 spot beams to cover the continental United States. This technology of using spot beams enables frequency reuse and allows for an increased number of local program offerings to DBS subscribers. The introduction of Ka-band satellites with spot beams and real-time traffic management information will increase local channel availability and enable implementation of broadband interactive TV.

Digital Audio Broadcasting

Satellite-delivered digital audio services are beginning to emerge that provide 100 or more channels of CD-quality radio programming in multiple languages to both fixed and mobile receivers. The technology of digital audio radio service (DARS) was first demonstrated commercially in 1999 by WorldSpace International Network, a Washington, D.C.-based company that launched AfriStar, the first of three planned geostationary satellites. AsiaStar was launched in early 2000 and is expected to become operational by year's end. AmeriStar is planned for launch during the first half of 2001. When all three WorldStar satellites are operational, they will provide digital audio broadcasting service to 80% of the world's population. WorldSpace's three GEO satellites are intended to provide service to the emerging and underserved markets of the world, including Africa, the Middle East, Asia, Latin America, South America, and the Caribbean. Each WorldSpace satellite has three regional beams, and each beam can broadcast more than 50 channels of crystal clear audio directly to palm-sized portable receivers. Programming options include channels for entertainment, news, information, education, sports, and culture. On AfriStar, for example, there are 23 broadcast services in 16 different languages. The first generation of WorldSpace satellites will provide only stationary reception via small receivers. The AsiaStar service, for example, will be broadcast to 4-inch dish antennas built into small radios. Future generation receivers will be designed for mobile reception.

In late 2000, two more DARS operators, Sirius Satellite Radio and XM Satellite Radio, plan to begin commercial operations to provide CD-quality music, news, and variety programming in the United States. Seeking to accelerate the growth of the DARS industry and to achieve the largest possible audience for their products, Sirius and XM Radio formed an alliance in 2000 to develop a uniform standard for satellite radios. New York-based Sirius Satellite Radio, formerly CD Radio, plans a three-satellite constellation. The first satellite, Sirius-1, was successfully launched in June 2000. Sirius-2 is slated for launching in September 2000, and Sirius-3 in October 2000. All three Sirius satellites were designed specifically for satellite radio broadcasting and will be among the first in the world to use the S band to deliver audio content. Sirius intends to have 100 channels operational by year-end 2000.

The Sirius satellites will be placed in inclined elliptical orbits—rather than GEO orbits over the equator—to maximize the line of sight to the satellites. The elliptical path will ensure that each satellite spends about 16 hours a day north of the equator and that two satellites are right over the United States all the time. These orbits enable the satellites to relay the Sirius broadcast signal to the United States from a much higher angle than GEO satellites. This will result in improved signal strength and coast-to-coast coverage. Content will be fed simultaneously to several transmitters in major urban areas. These terrestrial repeaters supplement satellite coverage in urban areas, where tall buildings may block the satellite signal. The Sirius satellite system is designed to broadcast as many as 100 music, news, information, and entertainment programming channels to motorists throughout the continental United States. Inside the car, drivers will find a digital display showing, for instance, the song's title, artist, record label, and running time. As with XM Satellite Radio as well, listeners driving from coast to coast will be able to stay on their favorite channels all the way. Ford Motor Company, BMW, and Daimler Chrysler all have agreements in place with Sirius to install new Sirius/XM-ready radios in new cars as standard equipment. Space Systems Loral built the three Sirius orbiting satellite plus one ground spare.

Similar to Sirius is XM Satellite Radio, which will also use terrestrial repeaters and will operate in the S band. XM's first satellite is scheduled to be launched in November 2000, and commercial service is set to begin during the first half of 2001. XM Satellite Radio was founded in 1992 and has its headquarters in Washington, D.C. Its major investors are American Mobile Satellite Corporation (now called "MOTIENT"), General Motors, DIRECTV, and Clear Channel Communications. XM Satellite Radio plans to broadcast as many as 100 brand-new radio channels to consumers in their cars, homes, and offices using two high-power Hughes HS 702 geostationary satellites and very small antennas. Listeners using portable and fixed home radios will be served by built-in antennas that are 2.4 inches in diameter. Programming is scheduled to include music, news, talk, sports, entertainment, ethnic, and children's channels, 24 hours each day. XM estimates its potential market at 60 million subscribers, who will pay about US \$10 per month. XM has already contracted with General Motors to place XM satellite receivers in all GM cars and trucks. The XM satellites have a communications payload built by Alcatel in partnership with Hughes Electronics; they will be operated by Telecast Canada. The future generation of XM satellites will deliver messages, transmit E-mail, and provide E-commerce applications to mobile platforms, thereby supplementing and augmenting other mobile services. In addition, simple two-way connectivity using the XM spectrum may be possible.

VSAT Networks and Digital Data

In the developed world, VSAT technology is widely used to establish quick, reliable, and cost-effective two-way data, voice, and video distribution networks. Networks can be citywide or worldwide, regional or domestic. Typically, they are used by large organizations that want to communicate in real time with their

local or regional branch offices and geographically dispersed customers, suppliers, and employees. Among the world's leading automakers—Ford, Chrysler, General Motors, Toyota, Fiat, Saab, Volkswagen, and Peugeot—all have proprietary VSAT networks. Other major VSAT networks include those operated by the Bank of China, Federal Express, Hewlett-Packard, IBM, Texas Instruments, Xerox, Home Depot, and the China People's Daily newspaper. VSAT networks perform a myriad of functions. These include electronic funds transfer; credit card authorizations at the point of sale, including "pay at the pump" gas stations; inventory control; ordering parts and products from a company's distribution center; providing corporate headquarters with up-to-the-minute sales data from all outlets; and telecasting corporatewide announcements from the organization's key executives at headquarters. There are currently thousands of VSAT networks in operation. Most were installed by one of three major competitors: Hughes Network Systems, Scientific-Atlanta, and Gilat Satellite Networks.

The use of VSAT networks is also very important for distance learning in both the developed and developing regions of the world. Distance learning via VSAT networks is used extensively in the developed world primarily for corporate training. A geographically dispersed sales force no longer has to fly to corporate headquarters to learn about a new product, for example. And auto service technicians can receive instruction from experts at major manufacturing centers without ever leaving their local dealerships. Through interactivity, questions can be asked from any remote site and relayed instantaneously to all remote sites and to the presenter at the originating location. In addition, VSAT networks enable corporate employees to earn an academic degree from a university across town or across the continent. The return audio path is used to ask or answer questions, and the return data path is used to take examinations. Universities, such as the National Technological University in Colorado, now exist specifically to provide instructional programming via VSAT and other types of distance learning networks.

Using VSAT networks for distance learning may be most valuable when they are located in less populated regions of the world, where children and adults can be taught to read, write, speak their national language, and practice a trade. Indonesia was one of the first countries to use a VSAT network for distance learning. Its citizens, spread out over more than 10,000 islands, continue to have the opportunity to learn the national language, Bahasa Indonesia, by VSAT. In the less developed world, VSAT networks can also play a vital role in health care. A patient's chart and X-ray films can be transmitted to a consulting specialist, who may issue a diagnosis from halfway around the world. Special medicines can be ordered instantly. And programming about good nutrition and proper dental care can help teach preventive medicine. Recently, the U.S. TV programmer, Discovery Communications, began funding VSAT-based satellite education in developing countries, providing advanced technological resources and training to rural and disadvantaged schools and community centers around the world. Discovery's first efforts are focused on sub-Saharan Africa and Latin America (24).

The worldwide VSAT market will continue to increase—until, gradually in the coming years—it is superseded by the ultra-high-speed broadband networks expected to debut in the 2002–2005 time frame. In the meantime, to manage the

increasing demand for fast two-way Internet links, driven in large part by the exploding growth of business-to-business E-commerce, companies such as Hughes Network Systems and EchoStar Communications Corporation are coming up with satellite solutions. In fact, more than 11% of the world's Internet Service Providers (ISPs) now use satellite links to connect to the Internet backbone. In 1999, the total value of the "Internet Protocol (IP) Over Satellite" market was US \$210.4 million. By 2000, that value had risen to US \$710.9 million. Increasing use of two-way satellite terminals to provide high-speed Internet access among intra- and intercorporate sites allows network owners to bypass telephone and cable companies and maintain end-to-end control over their networks (25).

As stated in the introduction to this article, Hughes Electronics plans to roll out an ultra-high-speed broadband satellite system in 2003. But Hughes will not have to attract Spaceway subscribers from scratch. In the intervening years, Hughes is offering DirecPC, a two-way satellite-based broadband Internet delivery system that delivers Web pages and software programs to home and small office satellite dishes at the rate of 400 kilobits per second. For downloads, that is faster than an ordinary telephone line but slower than a cable modem. Upstream requests, once handled by ordinary telephone, can now travel by satellite at a rate that is three times faster than a dedicated ISDN line. When customers of the "old DirecPC" are switched over to Spaceway, data will be delivered at up to six megabits a second. This is faster than today's cable modem delivery. Spaceway also promises an uplink capability of two megabits per second. In June 1999, America Online said it was investing US \$1.5 billion in Hughes Electronics to partner with Hughes to offer AOL's new broadband service, AOL Plus, over DirecPC.

In the fourth quarter of 2000, Hughes Network Systems intends to supercharge DirecPC with two-way satellite capability. This is necessary to allow consumers to downstream movies, games, and concerts into the home over the personal computer and to allow businesses to hold real-time videoconferences without the slow motion, blurry, and jittery video seen on today's VSAT networks. Offering "always on" capability, the new two-way high-speed satellite service will allow consumers to bypass the dial-up telephone or cable system completely and avoid land-based bottlenecks on the Internet. The two-way satellite version of DirecPC will operate on the current medium-power Ku-band satellites operated by PanAmSat, which is 81% owned by Hughes Electronics. This early entry two-way satellite service will be offered to enterprises around the world. For example, S Kumars.com, an Internet kiosk operator in India, has bought 50,000 DirecPC terminals for use throughout India.

Satellite Broadband Communications

According to the Teal Group, satellite industry consultants, some 384 satellites dedicated to broadband multimedia will be launched between 2000 and 2008. This constitutes 38% of all satellites launched during this period. These new Ka- and Ku-band satellites will help create a market for new GEO and LEO broadband satellite services worth more than US \$9 billion annually by 2004 (26). As stated before, advanced satellite broadband communication systems provide

significantly more bandwidth and much higher data bit rates than the broadband and narrowband systems currently in service. Broadband technology also allows multitasking of applications, that is, satellite-based broadband communication systems can handle numerous applications at the same time, for example, a live, full-motion digital videoconference and data exchange.

Broadband satellite users will be able to accomplish a variety of digital interchanges ranging from data to voice to video and multimedia. They will be able to download entire CD-ROM databases in a fraction of the time it currently takes. For example, in 2000, it takes 9 minutes on a 28.8-kbps phone line to download the Sunday edition of the *Washington Post*. By using a 1.5 Mbps satellite link instead, the time is reduced to about 10.4 seconds. The interactive multimedia capabilities of these systems will dramatically change and speed up the way we work. For example, broadband satellite users will have simultaneous access to high-data-rate on-line computer networking and low-cost, high-resolution interactive desktop videoconferencing. This means that an architect in China, for example, can work via broadband technology with one or more colleagues anywhere in the world, exchanging CAD/CAM images instantaneously. Currently, 2 megabits of CAD/CAM content can take 70 seconds to transmit on conventional phone lines. A broadband satellite link will reduce that time to just 1.4 seconds.

Remote manufacturing and control will be another workplace application for satellite broadband networks. In today's global economy, goods are often designed, manufactured, and marketed in different distant locations, and raw materials and consumers are located in still other venues. A broadband remote manufacturing/control system will allow an electronic toy designer in Seattle, for instance, to send specifications for a new toy directly to a manufacturing plant in Korea. In turn, the manufacturer could order the necessary microchips on a timely basis. The manufacturer also could use the same satellite broadband network to send production status data to the worldwide distribution network to allow for just-in-time delivery and better inventory control. Finally, the manufacturer, together with the distributor, could keep marketing and advertising agencies apprised of exact product delivery dates in different regions of the nation or the world.

As chip prices fall and compression rates soar, more and more companies are announcing plans for global satellite broadband systems and services. According to Pioneer Consulting, total global broadband revenues will increase from around US \$200 million in 1999 to US \$37 billion in 2008 (27). Virtually all international satellite operators, such as PanAmSat, EUTELSAT, INMARSAT, Telesat (Canada), and New Skies (the commercial spinoff from INTELSAT) already have plans to acquire broadband satellites. INMARSAT (the International Maritime Satellite Organization), for example, has announced the purchase of three global mobile broadband satellites for its new Broadband Global Area Network, which is to become operational during 2004. The INMARSAT satellites will be 100 times more powerful than INMARSAT's current world-leading global mobile 64-kbit/s network. In addition, the Broadband Global Area Network (B-GAN) will provide at least 10 times more capacity for new users, enabling INMARSAT to fulfill the growing need of global corporate enterprises for on-line high-speed access to information and communications. INMARSAT projects high

demand for mobile broadband services, and its satellites will also provide a seamless extension to fixed networks. According to a year 2000 forecast from ING Barings, the mobile satellite market will be worth more than US \$4 billion in 2004 and will double to more than US \$8 billion in 2009.

It is forecast that business-to-business E-commerce will be a \$US1 trillion market by 2003 (28). Six companies—Astrolink, EuroSkyWay, iSky, SkyBridge, Spaceway, and Teledesic—have announced plans for new satellite broadband networks, which they believe will give them a piece of this lucrative new market. Teledesic was one of the first to announce such plans, a US \$10 billion project funded by cellular pioneer Craig McCaw, Microsoft Chairman Bill Gates, Saudi Prince Alwaleed Bin Talal, the Abu Dhabi Investment Company, Motorola, and Boeing. Teledesic is building what it calls a global broadband “Internet-in-the-Sky” that is scheduled to become operational in late 2003. In addition to providing advanced business services, Teledesic aims to accelerate the spread of knowledge throughout the world and facilitate improvements in global education, health care, and the environment. The ambitious system will comprise 288 low Earth orbit (LEO) operational satellites, plus in-orbit spares, operating in the uncrowded, high-frequency Ka band of the radio spectrum. (In this frequency band, it is possible to achieve compact antenna terminals half the size of traditional VSAT Ku-band systems.) To use the radio spectrum efficiently, the Teledesic system will allocate frequencies dynamically and reuse them many times within each satellite footprint. The 288 LEO satellites will be divided into 12 planes; each will have 24 satellites orbiting at 1375 km (854 mi) above Earth.

Teledesic chose a low orbit to eliminate the signal delay that can be experienced in communications through traditional geostationary satellites and that could seriously impair interactive communications. A low orbit also enables using small, low-power terminals and antennas. The laptop-size terminals will mount flat on a rooftop and connect inside to a computer network or personal computer. Ground-based gateways will enable service providers to offer seamless links to wire-line and wireless networks, although fewer terrestrial facilities will be needed because all 288 satellites will be interconnected via intersatellite crosslinks. This reduces the time it takes to establish a connection. The Teledesic network is designed so that from anywhere on Earth, a Teledesic ground terminal can always “see” a satellite nearly directly overhead and without obstruction. This is accomplished by having an elevation angle of 40° or higher at all times in all locations. A lower elevation angle would increase the likelihood of obstruction by surrounding buildings, trees, hills, or other topographic features.

Most Teledesic users will have two-way connections that provide up to 64 Mbps on the downlink and up to 2 Mbps on the uplink. Sixty-four Mbps represents access speeds that are more than 2000 times faster than those available via today’s standard analog modems. A key technical challenge for Teledesic will be to develop low-cost antennae that can track moving satellites. Unlike narrow-band mobile satellite terminals, which can use simple omnidirectional antennae, broadband Teledesic antennae must be continuously repositioned so that they point to the correct satellite. Like all of the other satellite broadband systems announced, Teledesic will offer “bandwidth on demand.” This means that users pay only for the bandwidth they actually use and do not have to reserve bandwidth in advance or agree to a minimum monthly or annual service agreement.

For the network operator, bandwidth on demand means that the network can support a much higher number of users because capacity is released only when needed. The Teledesic network will cover nearly 100% of Earth's population and 95% of the landmass. It is designed to support millions of simultaneous users. Teledesic, founded in 1990, is based in Bellevue, Washington. In mid-2000, Teledesic was behind in its deployment schedule, and it was rumored that the company was considering reducing the number of satellites to be deployed.

By comparison, the EuroSkyWay global satellite broadband system will use a cluster of five operational satellites, all in geostationary orbit. It will offer an aggregate capacity of 45 Gbps, which will be available either on an on-demand or reserved basis. EuroSkyWay's Ka-band satellites will feature digital onboard processing, which has already been successfully tested in orbit since 1991 aboard two Italsat satellites built by EuroSkyWay's founding company, Alenia Aerospazio. In fact, Alenia Aerospazio has been developing payloads and satellites using the Ka band and onboard processors since 1984 as the prime contractor of the European Space Agency. Alenia Aerospazio developed the first commercial multimedia onboard processor in the world. It has been operating in orbit since March 1998 onboard the Eutelsat Hot Bird 4 satellite.

Like Teledesic, EuroSkyWay's satellites will also be interlinked. Using digital onboard processing and intersatellite links, EuroSkyWay satellites will provide full connectivity between any pair of spot beams. The satellites will also handle any packet-type switching, such as that required for two-way Internet communications. To overcome the so-called latency problem generated by the use of the Internet protocol TCP/IP for GEO satellites, EuroSkyWay will use a proprietary system based on the asynchronous transfer mode (ATM) protocol. In the first of two projected operational phases, EuroSkyWay will provide coverage across Europe, the Middle East, Greece, Turkey, Africa, and some former USSR countries. The first satellite is planned for launching by the year 2001. In the second and final operational phase, three additional satellites will provide increased capacity plus coverage across Africa and Asia. EuroSkyWay was founded in Italy in March 1997 by Alenia Aerospazio Space Division, the satellite manufacturing company of the Finmeccanica Group.

In addition to providing fixed services, EuroSkyWay will also bring broadband services to mobile users by offering communication links at a speed that is up to 200 times faster than traditional cellular phones. The company identifies its key potential customers as large telecommunications companies and Internet services providers, large multinational corporations, and social service users. In this latter category, EuroSkyWay specifies hospitals, research centers, universities, and all users that require multimedia interactive connections for social needs. The provision of medical services is a special directive. The company's slogan is "move information, not patients." In fact, using existing Alenia Aerospazio satellite facilities, technical support has been provided in such areas as Sarajevo, Bosnia-Herzegovina, and Kosovo. Doctors at medical institutes in Milan and Rome assisted the injured via satellite-based telemedicine.

As mentioned, Teledesic intends to fly an all-LEO system, and EuroSkyWay an all-GEO system. The global Ka-band Spaceway system from Hughes Electronics Company will consist of eight GEO satellites and additional satellites that will operate in lower Earth orbits. Ground stations will range from user

terminals with antennas approximately 26 in. in diameter to larger gateways for connectivity to terrestrial backbone networks. The initial Spaceway GEO satellite system will provide ubiquitous coverage in four main regions of the world. North America will be first to be brought on line, as early as 2002–2003 with two Hughes-built HS702 geosynchronous satellites plus an in-orbit spare. Hughes plans to work with global strategic partners to extend the system into other regions as markets develop, including Europe, the Middle East and Africa, Latin America, and Asia. Once this initial system is operating and producing sufficient revenue, the complementary non-GEO system will be introduced. It will expand the network capability to offer additional broadband and interactive multimedia services in high-traffic markets. Spaceway will extend the reach of traditional VSAT networks. It will seamlessly integrate with existing land-based systems and will be fully compatible with a wide range of terrestrial transmission standards. Spaceway will allow corporations to consolidate their local and wide area networks into single high-speed networks. Uplink rates will be between 16 Kbps and 6 Mbps. The satellite system will employ onboard digital processing, packet switching, frequency reuse, and spot beam technology, and will offer mesh connectivity throughout the service area. This connectivity, for example, will allow customers to communicate directly via satellite with other customers without having to go through a terrestrial retransmission hub. It also permits direct, full broadcast capability throughout the service area. Hughes and its subsidiaries own and operate the largest privately owned fleet of commercial satellites in the world. The company is also the world leader in providing satellite-based private business networks. The capabilities of these networks will be greatly expanded through Spaceway.

Thus far, all of the promised broadband systems mentioned will operate in the relatively unused Ka band, where there is a large quantity of frequency to be shared among several companies desiring to offer bandwidth- and frequency-intensive services. The final broadband system to be discussed here will not use the Ka band, but rather the Ku band—a set of frequencies used primarily for satellite news gathering, network transmissions to broadcast and cable affiliates, corporate videoconferences, and satellite-to-home broadcasting. Beginning in 2003, SkyBridge GP, Inc. plans to operate a US \$6.1 billion Ku-band LEO satellite system that will provide end users access to high data rate multimedia services. Like the other Ka-band services planned, SkyBridge wants to take advantage of the fast growing market demand for high-speed Internet access, corporate intra- and extranets, LAN/WAN remote access, E-commerce, videoconferencing, and interactive video and audio entertainment. The SkyBridge service is aimed at both residential and business users. According to SkyBridge, up to 72 million residential users will seek to purchase broadband communications services by 2005, and by the same year, businesses will spend more than US \$100 billion per year on broadband services. The 80 LEO satellite constellation from SkyBridge is designed to limit use of terrestrial facilities—except where these are more cost-efficient—and thus solve the “local loop” or “last mile” problem by using satellite antennas that connect directly to satellites.

According to SkyBridge, it will offer telecommunications operators a flexible, timely, and cost-effective alternative to slow terrestrial roll-out programs, for example, cable, fiber. The 80 Ku-band satellites will orbit at an altitude of

913 mi (1469 km). This low Earth orbit allows the short signal propagation time—30 milliseconds—needed to provide real-time interactive services. The SkyBridge system will comprise about 200 gateway Earth stations for worldwide coverage. Each gateway will have a 234-mile (350-km) radius of coverage. The gateway stations will interface with all existing terrestrial facilities through an ATM switch that will ensure seamless integration with residential or business satellite terminals that will cost about US \$700.

The SkyBridge system is designed to operate in the 10 to 18-GHz Ku-frequency band without causing interference to either GEO satellite operators or terrestrial users. Frequency will be shared between SkyBridge and the many different Ku-band GEO satellite operators by limiting the power from the SkyBridge system into the GEO transmit arc. Specifically, each SkyBridge satellite will cease transmissions during all potential interference conditions. A SkyBridge satellite will stop transmissions to a gateway cell, and the SkyBridge Earth stations in the gateway cell (including all end-user terminals) will cease transmission to the satellite when the satellite enters that gateway's "nonoperating zone." This zone will span 10° on either side of the GEO arc as seen by any Earth station in the gateway cell. The traffic in that cell will be handed over to another satellite in the constellation. The SkyBridge satellite system will have neither onboard switching nor intersatellite links. The gateways and satellites will be preprogrammed to hand over traffic.

As do the other Ka-band broadband operators, SkyBridge intends to serve both developed and developing countries around the world with its ubiquitous "instant" infrastructure. As is the case with Hughes's Spaceway system, many international partners are being recruited to help pay for these expensive systems. Alcatel, a multinational telecommunications firm located in Paris, is the general partner of SkyBridge LP. To assist in the design, development, and manufacture of its system, SkyBridge has arranged a diversified international consortium. Participants include Boeing (U.S.), COM DEV International (Canada), CNES (Centre National d'Etudes Spatiales) (France), EMS Technologies (U.S.), Litton Industries (U.S.), Loral Space and Communications Company (U.S.), Mitsubishi Electric Corporation (Japan), Sharp Corporation (Japan), SNECMA (France), Toshiba Corporation (Japan), a Belgian banking institution, and others.

Satellite Handheld Mobile Telephony

According to the Teal Group, some 450 new satellites for mobile telephony will be launched between 2000 and 2008. This represents 44%—nearly half—of all new satellites to be launched during the coming years (29). [Teal Group, report titled "Commercial Communications Satellites, Satellite & Launch Services Market Forecast: 1999–2008," April 1999] According to many, data delivery by phone will be the next boom technology for consumers. Using their handheld satellite phone receivers, they will be able to receive E-mail, access the Internet, obtain the latest news and weather forecasts, trade stocks, and participate in videoconferences via a built-in TV screen.

As mentioned before, INMARSAT, the International Maritime Satellite Organization, which operates a global constellation of geosynchronous satellites,

pioneered a worldwide infrastructure for mobile telephony and data communications as early as 1979. Today, new regional and global mobile constellations of GEO, LEO, and MEO (medium Earth orbit) satellites are coming on-line to compete with INMARSAT. These include global mobile systems such as Globalstar and INMARSAT's own privatized spinoff, New ICO Global Communications. New regional systems include MOTIENT (formerly American Mobile Satellite Communications, or AMSC; Thuraya; Asia Cellular Satellite Communications (ACeS); and Asia Pacific Mobile Telecommunications (APMT)). All of these services are aimed at the individual who needs uninterrupted global phone, fax, data, or pager access. This can include the business traveler in remote areas of the world or in developing countries where the telecommunications infrastructure is inadequate or nonexistent. Other users include workers involved in oil and gas extraction, those providing disaster relief, and those working aboard commercial and government ships.

The first LEO satellite mobile telephony system to become operational, Iridium, began commercial service in November 1998. Although Iridium—that has 19 strategic partners from around the world—was forced to terminate service in March 2000 for financial reasons, the system did prove the technology. The Motorola-designed system required US \$5 billion and 11 years to create and was authorized to provide voice, fax, and data transmission services in more than 120 countries. Iridium's 66 satellites, each weighing just 700 kg, formed a cross-linked grid only 485 mi (780 km) above the Earth. The Iridium satellites could receive the signals of a handheld phone. They acted like cellular towers in the sky, where wireless signals could move overhead instead of through ground-based cells. The Iridium network integrated land-based phone lines, local cellular facilities, and satellites to provide the quickest and most cost-efficient call routing. The Iridium handheld telephone worked in two modes. As a mobile cellular phone, it sought out available service from existing land-based networks. In this mode, it operated in the same way as today's terrestrial cellular systems. When cellular service was not available, the Iridium phone would switch to satellite mode. At least one Iridium satellite was always available overhead to receive a transmission. The call was then relayed from satellite to satellite, until it reached its destination, either through a local Iridium gateway and the PSTN or directly from the satellite to a receiving phone.

According to industry analysts, Iridium's market failed to materialize for five reasons. First, the handsets were large and heavy; second, they cost US \$3000 apiece; third, they were not functional indoors; fourth, Iridium chose to build its entire system up-front; it cost US \$5 billion and resulted in unsustainably high consumer rates; and fifth, it took too long (11 years) to get to market, so that initial market assumptions and technologies became outdated. For example, by the time Iridium was launched, cellular and personal communications service (PCS) technology had improved dramatically and was so pervasive that for most people, a traditional cellular phone was the optimum choice. Its lower cost and weight deterred conversion to the Iridium service.

Globalstar, another global mobile satellite telephony service, is now commercially available in more than 30 countries (as of May 2000) in North America, South America, Central America, Latin America, Asia, and Europe. Globalstar has said that it expects to offer commercial service in at least 50 countries,

including Russia and South Africa, by the end of June 2000. The company, a consortium led by Loral Space & Communications and Qualcomm of the United States, plans to provide service in more than 100 countries on six continents. Like its erstwhile competitor Iridium, Globalstar uses satellites to extend traditional mobile and fixed phone services and offers the flexibility to make both satellite and cellular calls through one phone anywhere in the world. Also like Iridium, Globalstar supports data transmissions and Internet connectivity as part of an overall range of services. E-mail messages are received through the system at rates of 9600 kbps.

Unlike the satellites in the Iridium network, Globalstar's 48 LEO satellites are not cross-linked and perform no onboard processing. Also, users are not connected directly to or from a satellite. Instead, Globalstar is what is called a "bent-pipe" system. Calls are routed via a terrestrial path from a fixed phone or handheld or vehicular mobile phone to a gateway Earth Station, which uplinks the call to a satellite. The call is then routed by satellite down to another gateway Earth station and then transmitted through local terrestrial wire-line and wireless systems to the final destination. The Globalstar system can pick up signals from more than 80% of Earth's surface. In fact, several satellites pick up a call (via the terrestrial gateway), and this so-called "path diversity" ensures that the call does not get dropped, even if one satellite moves out of sight of the phone. Each Globalstar satellite consists of an antenna, a trapezoidal body, two solar arrays, and a magnetometer. The Globalstar satellites were manufactured by Space Systems Loral.

The satellites operate at an altitude of 1414 km (876 mi), so that there is no perceptible voice delay or echo effect. The satellites are placed in eight orbital planes of six satellites each, inclined at 52° to provide service on Earth from 70° North latitude to 70° South latitude. The Globalstar system uses Qualcomm's CDMA (code division multiple access) transmission technology, that offers better voice quality and security than terrestrial cellular networks. Small, lightweight multimode handsets are used for either satellite or terrestrial cellular service (e.g., when inside buildings) in such standards as GSM, AMPS, and CDMA. Placing a call takes roughly 10 milliseconds. In addition to offering mobile telephony, Globalstar also intends to serve users in underdeveloped parts of the world with affordable fixed-site telephones. For example, these could be located in coin-operated village telephone booths or small business offices that currently cannot enter the global marketplace, for example, a supplier of batik in Indonesia that needs to communicate regularly with a wholesaler in New York. In the first quarter of 2000, Globalstar began service in China. The company plans to bring communications services to vast reaches of the country outside the range of existing cellular and wire-line phone systems. Globalstar estimates that 50% of the world has little or no phone service and believes that satellite telephony eliminates the high cost and long times needed to build land-line or wireless network infrastructures. Beside Loral and Qualcomm, the Globalstar consortium's strategic partners include Alenia Aerospazio (Italy), China Telecom (Hong Kong), DACOM (Korea), DaimlerChrysler Aerospace (Germany), Chrysler Aerospace (U.S.), Hyundai (Korea), France Telecom, Alcatel (France), Elsacom (Finland), and Vodaphone AirTouch (U.K.). As of February 2000, all 52 Globalstar satellites (48 operational and four in-orbit spares) were in orbit. However, as of

August 2000, subscriber numbers lagged far behind projections, and Globalstar's fate was in question. Despite the relative technological simplicity of the Globalstar network (compared with Iridium), the total system cost between US \$2 billion and US \$3 billion.

Beginning in the first quarter of 2001, another satellite global mobile telephony service will be provided by London-based New ICO Global Communications—so named after ICO Global Communications emerged from bankruptcy protection in May 2000. Like Globalstar, the New ICO system also maximizes use of terrestrial wire-line and wireless (cellular and PCS) networks, and the satellites are not cross-linked. However, New ICO users do have the option to uplink their calls to the satellite directly from the handset. When terrestrial transmission is selected, a dozen satellite gateway Earth stations around the world will provide access to and from the satellite. Using digital onboard processing, the New ICO system can handle 4500 simultaneous phone calls per satellite. The New ICO space segment will comprise 10 satellites operating in medium Earth orbit at an altitude of 6430 mi (10,355 km) and inclined 45° to the equator. The 10 operational satellites will provide complete, continuous overlapping coverage of Earth's surface. Like Globalstar, the New ICO system will employ the same path diversity, where a caller has access to more than one satellite at a time. New ICO believes that its MEO system has at least two advantages over LEO systems like Iridium and Globalstar. First, because there are far fewer satellites to build and launch, the capital cost of the New ICO system is lower, and these savings can be passed on to the consumer. Second, a higher satellite orbit means fewer call handoffs from one satellite to another and thus superior quality of service. The call handoffs are minimized because of a MEO satellite's larger footprint across the ground, where the user is, and because these spacecraft travel more slowly than LEO spacecraft, allowing the user handset more time to "see" the satellite.

New ICO's dual-mode mobile phones will be similar in appearance, size, and weight to standard GSM phones used throughout Europe. New ICO will also supply fixed telephones for such places as remote offices and oil rigs and coin-operated telephones for community phone booths. New ICO service is directed toward satisfying the needs of five market segments: (1) maritime; (2) remote fixed, for example, business and residential users located in areas where there are no existing fixed telecommunications services; (3) handheld mobile users who have an ongoing need for telephony where existing infrastructure is insufficient, as well as users who wish to extend the area in which they can roam; (4) transportation, for example, land transport operators whose needs include vehicle tracking and fleet management applications; and (5) government, for example, military and disaster relief agencies.

The launch of New ICO's first satellite onboard a Sea Launch rocket in March 2000 was unsuccessful. However, New ICO mitigated the impact of this failure with its plans to build and launch a total of 12 satellites even though the intended service requirements call only for 10 operational satellites in orbit. Because US \$4 billion were already spent, New ICO will need another US \$2.1 billion in financing to take it to commercial launch. New ICO Satellite Communications, based in London, was established in January 1995. The New ICO investor consortium comprises two of the world's leading telecommunications

entrepreneurs, Craig McCaw and Subhash Chandra, plus more than 60 of the world's leading telecommunications operators and manufactures. These include INMARSAT (U.K.), COMSAT (U.S.), ARABSAT (Saudi Arabia), British Telecom (U.K.), NEC Corporation (Japan), PT Indonesian Satellite Corporation, Singapore Telecommunications Ltd, Telecom Egypt, Telecomunicaciones de Mexico, Telefonica de Espana (Spain), Telekom Malaysia Berhad, Telstra Holdings (Australia), TRW (U.S.), Deutsche Telekom (Germany), Ericsson Ltd. (Sweden), Mitsubishi Wireless Communications (Japan), Telkom South Africa, Hughes Space & Communications Company (U.S.) that is building the satellite system, and Hughes Network Systems (U.S.).

Along with the global mobile satellite telephony systems described before, there will also be several regional systems in different parts of the world. Two, MOTIENT and Asia Cellular Satellite System (ACeS) are already operational. Two others expect to be operational in the 2000–2001 time frame. These are Thuraya Satellite Telecommunications and Asia-Pacific Mobile Telecommunications (APMT).

MOTIENT was founded in 1988 as AMSC and has been operational since 1996, when its first satellite, AMSC-1, was launched into orbit. The company targets U.S. corporations that have fleet management needs and services a footprint covering North and Central America, the Caribbean, and surrounding U.S. waters. Mobile workers can be linked to their companies for voice, data, dispatch, and messaging transmissions. Recently, mobile access to the Internet has been added. MOTIENT provides its services to several markets, including fixed-site satellite telephone users in remote areas; land mobile cellular/satellite telephone users; the maritime industry, including managers of multiple vessels and off-shore oil rigs; transportable regional dispatchers, for communications with their fleets of trucks, ships and other vessels, and railcars; and aeronautical users, for making and receiving satellite-linked telephone calls in flight on commercial and military planes. Calls are uplinked directly from the MOTIENT handset to AMSC-1 and downlinked to the MOTIENT gateway outside Washington, D.C., and from there through the PSTN. AMSC-1 can serve up to 10,000 mobile users simultaneously. Located at 101° West longitude, AMSC-1 operates in the L band and uses switchable spot beams to serve an ever changing, on-the-go market.

The main priority of ACeS, based in Jakarta, Indonesia, is to provide cost-effective fill-in satellite service for cellular operators and users and thereby afford seamless coverage throughout Southeast Asia, India, China, Australia, Indonesia, and the Philippines, among other countries. Like MOTIENT, ACeS will provide mobile and fixed voice services, data, fax, paging, and Internet access. The US \$900 million system uses two GEO satellites, one in orbit and operational and the second an in-orbit spare that will first serve as a backup and then later as an active satellite for system expansion into western and central Asia, eastern Europe, and parts of North Africa. The Garuda 1 satellite was successfully launched in February 2000 aboard a Proton/Block DM rocket from the Baikonur Cosmodrome in the Republic of Kazakhstan. After testing, it will become commercially available in the second half of 2000.

ACeS is the first regional, satellite-based, handheld mobile telecommunications system designed exclusively for the Asia-Pacific region. It is also the first integrated GSM (general services for mobile)/satellite network in the world.

Garuda 1 is said to be "the most powerful commercial GEO satellite ever built" and can cover more than 11 million square miles. Lockheed Martin Global Telecommunications is the manufacturer. Each Garuda satellite has 11,000 circuits. Because of the high power of Garuda I, the satellite can handle direct satellite uplinking and downlinking from any user's mobile phone handset. The 205-g Ericsson R190 dual-mode (GSM/Garuda) satellite phone is said to be the smallest satellite phone available worldwide.

The GSM cellular standard is operational in more than 100 countries worldwide. Subscribers to GSM networks worldwide can use ACeS while traveling in Asia by using an authorization card and a leased or owned ACeS satellite terminal. Once out of cellular range, the phone will automatically switch to satellite mode to send or receive calls and will switch back to GSM when in GSM coverage areas. Consumer retail pricing for use of the ACeS system will be less than US \$1 per minute, which is the lowest price for any of the satellite handheld mobile telephony systems announced. Service will be offered initially in eight licensed Asian countries whose total population is 1.7 billion—less than 11% have access to wireless service. These countries are Indonesia, the Philippines, Thailand, India, Pakistan, Bangladesh, Sri Lanka, and Taiwan. After initial service rollout, ACeS will make its network available to an additional 26 countries with which it has already signed international roaming agreements. Significantly, because of its vast population and inadequate telecommunications infrastructure, China is one of the countries that has signed a roaming agreement with ACeS. The four strategic partners of ACeS are Pasifik Satelit Nusantara (Indonesia); Philippines Long Distance Company; Jasmine International (Thailand); and Lockheed Martin Global Telecommunications (U.S.), which owns a 30% interest in ACeS.

Founded in 1997, Thuraya Satellite Telecommunications Company of the United Arab Emirates will serve an area that spans the Indian subcontinent, central Asia, the Middle East, north and central Africa, Turkey, and Europe. This area, much of which is sparsely populated, has rugged terrain, little telecommunications infrastructure, and comprises approximately 100 countries and two billion people. This will be the first and largest mobile satellite service in these regions. Thuraya plans to begin commercial service in late 2000 or early 2001 after a scheduled launch in September 2000. The Thuraya system is valued at US \$1 billion, including one in-orbit satellite at 44° East longitude and one ground spare. Thuraya will operate a high-power geosynchronous satellite to provide services that include the transmission of voice, data, fax, and messages, as well as location determination using the Global Positioning System of the U.S. Air Force. Thuraya will transmit and receive calls via a single 12.25-meter-aperture reflector on the satellite. Calls can be uplinked directly to the satellite from the user handset or relayed via a terrestrial path to a gateway satellite Earth station. The satellite employs onboard digital signal processing to create more than 200 spot beams that can be redirected in orbit, enabling Thuraya to adapt to consumer demand in real time. The system can handle 13,750 simultaneous voice circuits. Thuraya's investors include some of the leading telecommunications operators from the areas to be served, including Emirates Telecommunications Corporation (UAE); Arab Satellite Communications Organization (ARABSAT) (Saudi Arabia); Qatar Telecom; General Post and

Telecommunications Company (Libya); and other telecommunications organizations from Oman, Yemen, Egypt, Morocco, Sudan, Tunisia, and Germany. Another key investor is Hughes Space & Communications International, which built the Thuraya orbital and ground systems.

Finally, Asia Pacific Mobile Telecommunications Company (APMT), an early market leader, that has a satellite already in an advanced construction phase in 1998, experienced a major setback in early 1999 when the U.S. government withheld an export license from Hughes Space & Communications for the sale of the APMT satellite. (All but one of APMT's owners were Chinese companies, and on the news, Singapore Telecom dropped out.) APMT immediately cancelled the US \$450 million contract with Hughes and has had to start from scratch to seek other suppliers outside the United States. This will considerably delay service introduction. The APMT satellite was originally scheduled for launching in late 1999, and commercial operations were scheduled to begin in 2000. APMT's aim was to complement and extend terrestrial fixed and mobile coverage at low cost. Its coverage and pricing were going to be virtually the same as ACeS, which leaves APMT at a distinct competitive disadvantage at this time.

Regulation

Regulation and the globalization of satellite communications are both very important factors in determining the future development of satellite communications. Therefore, we present here a short description of the current situation with respect to these activities.

International Telecommunication Union. All communications entities that use or require radio-frequency spectrum are subject to the rules and regulations of the International Telecommunication Union, a specialized agency of the United Nations headquartered in Geneva, Switzerland. In the case of satellite communications, these regulated entities include operators of fixed (geostationary), broadcast (direct-to-home), and mobile satellite services, as well as operators of satellite ground links. Most, if not all, are also subject to national legislation and to the rules of numerous national telecommunication regulatory agencies (such as the Federal Communications Commission in the United States).

The mission of the ITU is to facilitate the global development of telecommunications for the universal benefit of mankind, through the rule of law, mutual consent, and cooperative action (30). The ITU comprises 189 country members and some 600 members from the private sector (as of July 2000). Among its primary goals is to create and oversee an orderly, transparent system to ensure satellite interconnection and interoperability of different national, regional, and global satellite systems on a technical and administrative basis.

In support of this goal, the ITU attempts to establish fair and equitable regulatory processes, sets standards, allocates orbital slots, provides for the efficient use of radio spectrum, supports telecommunications development worldwide, and ensures freedom from harmful interference through efficient spectrum management. It is safe to say that we could not have the robust satellite systems and applications we have today—and look for tomorrow—without the ITU's

equitable regulation of the world's limited radio-frequency spectrum and its enforcement of those regulations.

One of the most important tasks of the ITU today is to ensure that increasingly scarce radio spectrum is not allocated for uses that are not currently feasible or practical. The ITU is aware that it must strike a careful balance between its "first-come, first-served" policy of spectrum reservation (even where no satellite system has been developed) and accommodating deserving, but later, applicants whose well-engineered systems are ready to go (31).

The ITU is also charged with recommending tariff and accounting rate principles to ensure nondiscriminatory treatment. Members and spectrum users pay an annual membership fee to the ITU, along with various filing fees when applying to operate a new system, but they do not pay for the spectrum they use. All ITU decisions are reached by membership vote. Only government members vote on decisions regarding spectrum (32).

The origins of the ITU date back to 1865, when telegraph communications were heating up. There was no international regulatory body to take the lead in providing for the interconnection of national networks or to standardize equipment, issue operating instructions, and lay down common international tariffs. On May 17, 1865, the International Telegraph Union, later to become the International Telecommunication Union, was established (33).

ITU activity in satellite communications took off after the launch of Sputnik. In 1959, a study group was formed to study space radio communication. In 1963, an Extraordinary Administrative Conference for space communications was held to allocate frequencies to the various space-based services (33). Since that time, the ITU has held timely regional and world administrative conferences to plan for and provide new telecommunication services. Each major innovation in the field of telecommunications is matched by specific action on the part of the ITU to integrate new technologies into the world network and to provide the necessary resources (e.g., radio spectrum, orbital slots) to respond effectively to the expectations of member states. For example, with the advent of non geostationary low and medium Earth orbit satellite systems, the ITU assigns orbits rather than orbital slots. The orbit is defined by five factors: spacing of the satellites, altitude, inclination, number of planes, and number of satellites per plane.

The ITU's workload has become increasingly heavy and complex since the end of the Cold War. As the bipolar Cold War environment gave way to the global marketplace, satellites began to assume an increasingly important role in the communications infrastructures of the new CIS republics (34). This is also true of numerous developing nations, which at this time also began to seek a means of entry into the global marketplace. Satellites emerged as the quickest and most cost-efficient route to global telecommunications, creating "instant" ubiquitous telecommunications infrastructures. Through satellites, any country could enter the global marketplace immediately—without waiting years for installing copper wire or fiber connections.

The significant capital required for these satellite-based infrastructures made use of private funds inevitable. In most cases, the governments of the CIS and developing nations could not afford this infrastructure on their own, so private companies moved in. Moreover, the economic value of satellite slots and

radio spectrum and the provision of satellite services drew in other competing private companies. In fact, today, competition in providing satellite communication services is pervasive throughout the world, domestically, regionally, and globally.

For the ITU, this altered satellite communications landscape meant dozens of new members, new requirements for satellite orbital slots and finite spectrum, member demands for timely action despite growing complexity, and increased politicization in the face of mounting worldwide competition. Meanwhile, new regional and global satellite services increased competition for space throughout the orbital arc. Previously, most applicants sought only to provide a national service—often with a single satellite; now applicants that seek to provide regional and global satellite services want to be everywhere. Hence, they require orbital assignments above countries throughout the world (34).

In addition, new satellite technologies have emerged. These include interactive direct-to-home broadcasting, handheld mobile telephony, and high-data-rate broadband systems for interactive multimedia. These new technologies place added pressure on the ITU, which must ensure their technical soundness and allocate new orbital and spectrum resources so as to prevent harmful interference with any existing or planned system and speed their entry into the highly competitive world marketplace.

To promote the proliferation of global mobile telephony, for example, the ITU, in March 1999, set a single worldwide standard to enable all mobile phones to work anywhere in the world. Previously, global roaming was not feasible because there were competing and incompatible standards among nations. For instance, U.S. wireless networks used code division multiple access (CDMA), but Europe and large parts of Asia used GSM global system for mobile communication. In adopting a single standard, the ITU took the best from both technologies.

Along with the new satellite technologies, there are also competing terrestrial systems, such as fiber-optic broadband and cellular that must be accommodated with a spectrum.

The World Trade Organization. Hand in hand with privatization and competition in satellite communications has come deregulation, whereby many governments now allow foreign competitors to provide satellite services in once-closed markets. On January 1, 1998, the World Trade Organization's landmark agreement on basic telecommunications took effect, supporting the creation of a multilateral framework for trade in telecommunication services and open competitive markets. Coincidentally, the same date marked the opening up of telecommunications markets within most member states of the European Union. Both actions have a common objective, namely, to ensure nondiscriminatory market access for service providers and to lower costs for end users.

The telecommunication sector, including communication satellites, is one of the major components of the world's economy. The value of telecommunication sales (equipment and services) exceeded US \$1 trillion in 1998. Moreover, telecommunication networks are a major facilitator of trade in other goods and services, as well as in the timely movement of capital throughout world markets. Because of the importance of telecommunications, the WTO has undertaken to ensure that (1) there is free market access, (2) regulation is

transparent and nondiscriminatory, and (3) pricing is fair and nondiscriminatory (31).

Before 1998, only a handful of countries permitted competitive provision of international telecommunications services. Of the WTO's 188 member states, 69 have voluntarily agreed to comply with WTO agreements to liberalize telecommunications trade. These 69 signatories account for more than 90% of international telecommunications traffic. The WTO believes that a significant by-product of this reform will be quicker introduction of new technologies and services in signatory nations.

The WTO agreement on telecommunications underscores the fact that telecommunications trade has moved from bilateral to multilateral status. In large part, communication satellites are responsible for this significant change, because satellites do not recognize geographic borders. Moreover, the cost of satellite transmission is insensitive to distance. Telecommunications traffic that once moved from one country to another now flows through vast networks that can comprise dozens of countries. Where interconnection rates used to be settled by two nations, now many nations are involved, introducing greater complexity. Reform and fairness in assessing accounting rates is a primary objective of the WTO. Specifically, the WTO seeks a framework that is nondiscriminatory, cost-based, transparent, and consistent with the principle of voluntary multilateralism (31).

Conclusions

Looking back, at the end of 1999, the worldwide satellite industry had generated more than US \$69 billion in annual revenues. Communication satellite services, the largest and fastest growing segment of the industry, generated US \$30.7 billion in 1999. The U.S. satellite industry accounted for US \$31.9 billion of the total, or roughly 46% of worldwide revenues (35). This rather unbalanced spreadsheet will change quickly as more and more non-U.S. countries provide regional services, for example, satellite handheld mobile telephony and direct-to-home TV, and strengthen their manufactures' ability to compete better with U.S.-based rivals. As we have seen, satellite handheld mobile telephony and broadband services are already proceeding without U.S. involvement. The loss of APMT's US export license virtually invites international rivalry in manufacturing.

In August 2000, the mobile satellite services (MSS) market was the riskiest of all satellite businesses discussed here. Despite cumulative investments in excess of US \$20 billion and 25 years of effort, MSS remains a niche market with only an estimated 600,000 users worldwide. The current market is professional and draws subscribers primarily from areas where terrestrial cellular systems are not available. INMARSAT and Qualcomm's Omni Tracs service are the dominant players. Due to the failure of Iridium in early 2000, many market analysts doubt the viability of MSS as a successful business. Still, based on current market projections and growth rates, one can forecast the existence of three million to four million MSS users by 2004. Most of the growth will come from niche and specialized markets and from converting the telephony system to a medium-rate

data system. Such conversion has already taken place as Globalstar and New ICO begin to offer data service. In the future, all MSS operators hope to boost traffic and revenue with data applications. A significant mobile satellite data market is most likely to require specialized providers to address a wide range of asset tracking and telemetry applications.

Looking ahead at global broadband satellite services, Pioneer Consulting has forecast that total revenues will increase from around US \$200 million in 1999 to US \$37 billion in 2008. Moreover, according to the Teal Group, by 2008, annual revenues deriving from all commercial satellite technologies and services worldwide will total US \$120 billion. These skyrocketing revenues can be attributed to at least eight trends that have propelled the commercial satellite industry this far: (1) global deregulation and privatization, (2) digital transformation, (3) market convergence, (4) globalization of services, (5) strategic international alliances, (6) the dynamic emergence of—and rapidly accelerating global demand for—satellite broadband services, (7) worldwide demand for more mobile services, and (8) overall demand for more satellite-based telecommunications services of all kinds.

Global Deregulation and Privatization. Deregulation and privatization of satellite services has occurred throughout the world. This has led to the provision of commercial satellite services by entities from foreign countries that had previously been banned from providing these services in certain highly regulated countries, for example, India, China, Mexico, and Greece. The ensuing competition in turn has led to better quality products and services; lower, more competitive pricing; introduction of new product and service offerings; and spectacular growth in demand. Likewise, the decision of the U.S. Air Force to make its Global Positioning System available commercially to anyone in the world has triggered myriad new service offerings at competitive prices. Finally, when examining privatization, the case of INTELSAT and its commercial spin-off, New Skies, demonstrates the consumer benefits. Before PanAmSat had grown its satellite fleet enough to compete with INTELSAT, only INTELSAT provided satellite coverage around the world. As a monopoly, INTELSAT's prices were high, and new services were slow to come to market. Now, with New Skies operating a commercial satellite fleet that covers most of the world, competition (with both INTELSAT and PanAmSat and other satellite operators) has brought its familiar benefits. In addition, New Skies has demonstrated its commitment to innovation. Several different satellite services are provided on a single C/Ku-band satellite: DTH satellite TV broadcasting, point to multipoint distribution to cable headends, corporate data networks, Internet access, and multimedia transmissions.

Digital Transformation. Fully packed chips and very high rates of digital compression help give consumers the advantage of advanced satellite-based services at ever lower prices. Being able to install more and more information on chips, coupled with soaring digital compression rates, is making possible the multiuse satellite; without these technologies, satellites would be too heavy and expensive to launch. For example, Eutelsat's SESAT satellite, which was launched into GEO orbit in April 2000, has 18 Ku transponders that are used for a full range of services, including data and video broadcasting, direct-to-home TV, inter- and intracorporate networks, satellite newsgathering, Internet

backbone connections, high-speed Internet access, distance learning, high-speed transfer of software, and messaging and positioning services for mobile users. The satellite was dedicated to communications satellite pioneer Arthur C. Clarke and is so named.

Market Convergence. In the area of service provision, market convergence is giving consumers a whole new range of satellite products and services. For example, the strategic partnerships between TiVo and AOL TV with Hughes' DIRECTV merge the content suppliers with the satellite operator. This gives consumers a variety of advanced services, such as watching TV while accessing the Internet, writing E-mail, or requesting an advertiser's annual report. There is also a new, wide-ranging alliance under which News Corp. provides access to Yahoo! Web sites through its global satellite systems and allows Yahoo! to use News Corp.'s news and entertainment products, such as Fox News Channel and the *London Times* as a major source of content for its computer sites. Another example is the convergence of the automobile industry, for example, General Motors' OnStar system with PanAmSat's satellite fleet. What was initially an emergency system for obtaining assistance while driving has evolved into much more. Today, drivers of upscale General Motors cars (and the cars of some other manufacturers) can use the onboard satellite system to obtain directions to a selected destination, choose a restaurant or hotel and make reservations, or find the nearest gas station or dry cleaner. As described before, the convergence of manufacturing with satellite technology providers has revolutionized multinational manufacturing and distribution operations. Today, numerous industries and small businesses worldwide are finding new ways to take advantage of satellite technologies.

Globalization of Services. As deregulation and privatization occur, satellite-delivered services have become globalized to the extent that satellite TV viewers in southeast Asia have become ardent viewers of such fare as MTV, Nickelodeon, and "ER." Of course, CNN is virtually ubiquitous, available on almost any TV set in the world. But America is no longer exporting only its TV programs. A new service on DIRECTV allows American viewers to experience various cultures from around the world. WorldLink TV, which is telecast via DIRECTV, delivers public affairs programming from Belfast to Beijing. Viewers can tune in to independent newscasts, social documentaries, world music, and documentaries on human rights issues, the environment, and the global economy. Meanwhile, Discovery Communications of the United States is funding and providing satellite education in developing countries. Discovery is providing advanced technological resources and training to rural and disadvantaged schools and community centers around the world. In its first phase of operations, Discovery is establishing learning centers equipped with satellite and video technologies in sub-Saharan Africa and Latin America.

Strategic International Alliances. Strategic international alliances have become an economic necessity in the light of today's expensive satellite programs, including satellite construction, launch, insurance, and ground systems. This is increasingly the case as new satellite services are global, so that strategic alliances are formed for economic reasons and also for marketing purposes. For example, the ill-fated Iridium global satellite telephony service spent US \$5 billion to build its system and relied on 19 strategic investors from around the

world. One of its successors, Globalstar, has 10 founding partners and many additional financial supporters. The regional Thuraya satellite mobile telephony system has 15 strategic partners and investors.

Turning to broadband, to come on line, the emerging satellite broadband systems will need anywhere from less than US \$1 billion for iSky, which is using satellite capacity on Telesat (Canada), to US \$10 billion for Teledesic. Numerous strategic partnerships for both financing and marketing these broadband systems are essential. The stakes are high. For example, a team of WTO economists estimates that by year-end 2000, there will be more than 300 million Internet users worldwide and that E-commerce will equal US \$300 billion. Because most of the satellite broadband systems will not be on line until the 2001–2002 time frame, these numbers reflect the use of cable and other terrestrial technologies. Satellite broadband technology is certain to eliminate the bottleneck and expand the number of commercial consumer and business users. Laying the groundwork for successful satellite broadband companies will require strategic partnerships that enhance the program's technological, marketing, and economic success.

Soaring Demand for Satellite Broadband Services. Meanwhile, Pioneer Consulting forecasted that total global broadband satellite revenues will increase to US \$37 billion in 2008. Residential service will represent the majority of revenue gains, accounting for almost US \$22 billion of revenues in 2008. Demand for satellite broadband services will skyrocket as more and more business and residential consumers access the Internet. Netscape founder Marc Andreessen has forecast that, by 2010, the Internet will be 1000 times bigger than it is today and that satellites will be a key conduit because of their ubiquitous reach. China's Internet growth—even though stymied to some extent by government prohibitions—says a lot for growth around the world. In 1999, the number of China's Internet users climbed from 2.1 million to more than 6.7 million. By 2003, it is predicted that China will have more than 33 million Internet users and that the number will grow at an annual rate of 60% through 2004. As a whole, Asia had 40 million Internet users as of June 2000; quickly escalating growth is expected to bring the total to 375 million by 2005.

Demand for "On the Go" Satellite Mobile Services. It is estimated that, by 2003, more than one billion mobile phones will be in use around the world. Wall Street, wireless service providers, and industry analysts believe that wireless data will be the next area of explosive expansion in the Internet economy. Users of wireless phone systems, and particularly satellite-based handheld mobile phones, will use their phones for everything from learning about the local weather to videoconferencing across a continent. The handheld satellite phone will have a screen for displaying both information and broadcast images. Throughout the world, business commuters and travelers lose important business resources that they take for granted when they are at their desktop personal computers. These include Internet access, E-mail and faxing capability, and word processing. Many of these resources will soon become available via mobile platforms, on demand for users on the go. For example, INMARSAT's satellite-based B-GAN network, scheduled to become operational during 2004, will deliver Internet and intranet content, video-on-demand, videoconferencing, fax, E-mail, voice and LAN access speeds at up to 432 kbit/s virtually anywhere

in the world via notebook or palm top computers. According to INMARSAT President and CEO Michael Storey, the new system will enable INMARSAT to "meet the high demand for mobile broadband services and provide a seamless extension to fixed networks." INF Barings has forecast that the mobile satellite market will be worth more than US \$4 billion in 2004 and will double to more than US \$8 billion in 2009. INMARSAT expects to generate wholesale revenues from its new B-GAN network in excess of US \$10 billion, "as global multinational enterprises....have access to global mobile broadband communications wherever they are in the world," Storey added. Similarly, Boeing is prepared to afford airline passengers global mobile high-speed on-line connections, television, and other entertainment. Passengers will be able to use laptops to surf the Net at speeds up to 140 times faster than the fastest standard modem service. They will also be able to send and receive E-mail, transmit/receive word processing documents, and much more. Commercial cars with multimedia cockpits are already making their appearance. A personal computer hooked up to a swiveling satellite antenna to navigate the Internet gives access to the same broad array of information resources a driver has at the office. Passengers can even watch films.

The Importance of 24/7 Backup. Whether on the go, at home, or at the office, people have become accustomed to using or enjoying satellite-delivered services, and all indications show that their reliance on such services will increase rapidly in the future. Meanwhile, new subscribers are signing up by the thousands all over the world, from cosmopolitan cities to rural villages in undeveloped countries. The worldwide commercial satellite industry is worth hundreds of billions of dollars in terms of the revenue generated by satellite services. This revenue is significant both to the satellite industry as a whole and to national economies. And in between are businesses whose own revenues depend on the availability of satellite-based telecommunications.

In fact, the real value of commercial satellite-based telecommunications, information, and entertainment services is so diversified that it may not be possible to quantify it exactly. However, a new U.S. satellite company has introduced a product that underscores the vital importance of commercial satellite services for operators, users, and national economies alike. The company, called AssureSat, Inc., will provide two high-powered, specially designed geostationary satellites that will provide in-orbit backup protection service beginning in 2002. Once in orbit, the AssureSat satellites will be able to provide backup protection to most fixed GEO satellite operators by moving quickly to appropriate orbital slots to take over the communications tasks of malfunctioning spacecraft or spacecraft whose launches have failed. In-orbit operational backup can provide a cost-effective way for a satellite operator to retain customers and protect both its revenue stream and its customer relationships. The unique design of the spacecraft will allow them to operate on all three International Telecommunication Union (ITU) region frequency plans, and the antennas will be steerable in flight to offer variable footprint coverage in both the C-and Ku-frequency bands. Each of the AssureSat satellites will carry 36 C-band transponders and 36 Ku-band transponders. AssureSat will not offer backup protection for mobile satellites, thus leaving a significant new business opportunity for another entrepreneurial satellite company.

BIBLIOGRAPHY

1. *Worldwide Satellite Market Forecast: 1998–2007*. Teal Group, Fairfax, VA., 1999.
2. Los Angeles Times, Jan. 15, 2000, p. C 1.
3. Cahners In-Stat Group.
4. Annual EUTELSAT survey conducted by researchers Gfk, Gallup, and Nielsen, 1999.
5. Baskerville Communications Corporation.
6. Cahners In-Stat Group.
7. Paul Kagan Associates, Carmel, California, as cited in *Skyreport.com e-news*, 9/16/99.
8. *SkyReport*, September 16, 1999.
9. *www.astrolink.com*
10. TIME Magazine, February 28, 2000.
11. Internet Via Satellite Report 2000. DTT Consulting, Hampshire, U.K.
12. A system from Hughes Network Systems. *Forbes*, January 24, 2000.
13. *Satnews*, May 11, 2000.
14. *E-News@SkyReport.com*, April 19, 2000.
15. Study from Allied Business Intelligence, Oyster Bay, NY, May 2000.
16. *SatNews*, March 13, 2000.
17. Veteran former Hughes Electronics Satellite executive, Bruce Elbert.
18. *Wired News*, April 26, 1999.
19. MSNBC News Services and Associated Press, MSNBC.com, May 2, 2000.
20. *www.Astrolink*, The Marketplace, May 1, 2000.
21. Eutelsat study conducted by Gfk, Gallup, and Nielsen organizations.
22. *SatNews*, May 31, 2000.
23. Study, Cahners In-Stat Group, 10/21/99.
24. *Sky Report*, October 29, 1999.
25. Internet Via Satellite Report. DTT Consulting, Winchester, Hampshire, U.K., March 2000.
26. *World Broadband Satellite Service and Equipment Markets*. Frost & Sullivan, Mountain View, CA, 1998.
27. Pioneer Consulting. Next Generation Broadband Satellite Networks report. September 1999.
28. INMARSAT press release, INMARSAT to Deliver Mobile Global Broadband Data Communications. London, U.K., May 11, 2000.
29. Teal Group. Commercial Communications Satellites, Satellite & Launch Services Market Forecast: 1999–2008. April 1999.
30. *www.itu.ch*.
31. Interview with Michael Fitch, Vice President, Regulatory Affairs and Spectrum Management, Hughes Communications, Inc., January 26, 1999.
32. Jordan, P. Spectrum is gold. *Hughes Electron. Rev.* 4–5 (November/December 1998).
33. *www.itu.int/itudoc/about/itu/history*
34. The Revolution in International Telecommunications and the Role of the ITU, speech by Dr. Pekka Tarjanne, Secretary-General, ITU, January 16, 1998.
35. Satellite Industry Association and Futron. Third Annual Global Satellite Industry Indicators Survey. June 2000.

STEVEN D. DORFMAN
WAH L. LIM
Hughes Electronics Corporation
Los Angeles, California

COMMUNICATION SATELLITE DEVELOPMENT IN RUSSIA

Introduction

Work on the practical applications of artificial Earth satellites began in Russia (USSR) in 1961. The first of these satellites, the small low-altitude Strela series spacecraft of the “electronic mail” type, were developed by the Scientific Production Association for Applied Mechanics (NPO PM) and placed in orbit in 1964. One year later, the Experimental Design Bureau-1 (OKB-1), (currently S.P. Korolev Energia Rocket and Space Corporation) inserted the active relay satellite Molniya-1 into a highly elliptical orbit. In 1965, work on this spacecraft was transferred to NPO PM and, for the next 30 years, this organization was the only one in the country developing telecommunication satellites (Table 1). The first geostationary satellite was inserted into orbit in 1975.

It was not until 1997 that a communication satellite, Kupon, developed by another organization, the S.A. Lavochkin NPO, was placed in orbit, and in 1999, the communication satellite Yamal developed by Energia was launched. At present, work to design communication satellites is being conducted by the Machine Building NPO and the M.V. Khrunichev State Space Scientific Production Center.

The spacecraft that were and are being developed in Russia may be divided into three categories on the basis of the types of orbit:

- satellites in low circular orbits, primarily at an altitude of 1500 km (such satellites are used to form multisatellite communications systems);
- *Molniya* type satellites in highly elliptical orbit;
- satellites in geostationary orbit.

Satellites in Low Orbit

The first launch of a Strela type satellite took place on 22 August 1964. This was the first spacecraft to enable exchange of information among remote users. The satellites were launched together, five to eight at a time, into a circular near-polar orbit at an altitude of 1500 km. Each nonoriented small satellite had a mass of 50–80 kg and carried a retransmitter operating as part of an “electronic mail” system and a command and instrumentation system that transmitted telemetry data to Earth on the status of the satellite’s onboard systems and received control commands from Earth.

The satellites’ electric power supply came from a system consisting of solar cells and storage batteries. The solar cells provided energy for spacecraft equipment during the illuminated part of the orbit, during which the storage battery was also charged. In the dark part of the orbit, the equipment operated on power from the storage battery.

Table 1. The Spacecraft Developed by NPO PM and the Dates of Their First Launch

No.	Date	Satellite	Purpose(s)	Developer and manufacturer
1	8/22/64	Strela-1	Data relay	NPO PM
2	5/25/67	Molniya	Telephone, television broadcast	OKB-1 (manufactured by NPO PM)
3	7/2/69	Molniya-1B	Telephone, television broadcast	NPO PM
4	4/25/70	Strela-1M	Data relay	NPO PM
5	11/4/71	Molniya-2	Telephone, television broadcast	NPO PM
6	11/21/74	Molniya-3	Telephone, television broadcast	NPO PM
7	12/22/75	Raduga	Telephone, television broadcast	NPO PM
8	10/26/76	Ekran	Television broadcast	NPO PM
9	12/19/78	Gorizont	Telephone, television broadcast	NPO PM
10	12/17/81	Radio	Data relay	NPO PM
11	5/18/82	Potok	Data relay	NPO PM
12	4/22/83	Molniya-1T	Telephone	NPO-PM
13	10/25/85	Luch	Data, telephone, television broadcast	NPO PM
14	12/27/87	Ekran-M	Television broadcast	NPO PM
15	6/22/89	Raduga-1	Telephone	NPO PM (manufactured by Polet PO)
16	7/13/92	Gonets-D	Data relay	NPO PM
17	1/20/94	Gals	Telephone	NPO PM
18	10/13/94	Ekspress	Telephone, data relay	NPO PM
19	10/11/95	Luch-2	Data relay, telephone, television broadcast	NPO PM
20	2/19/96	Gonets-D1	Data relay	NPO PM
21	3/12/00	Ekspress-A	Telephone, television broadcast	NPO PM
22	4/18/00	Sesat	Telephone, television broadcast	NPO PM
23	7/20/01	Molniya-3K	Telephone	NPO-PM

The spacecraft equipment used only a small amount of power, and, thus, emitted only small amounts of heat. Thus, a passive thermal regulation system was used to maintain the thermal conditions required for equipment operation. These conditions were maintained by selecting the optical coefficients of the spacecraft and instrument exterior on the basis of the heat flux from the Sun and Earth.

The spacecraft design consisted of two assemblies, a pressurized container for the equipment and the exterior of the solar panels, which was approximately spherical in shape.

The Gonets type of satellite marked a further development of data relay systems. These satellites were also launched in groups into circular orbit of an altitude of 1500 km and an inclination of 82.5° . The Gonets-D satellite was designed to relay data in digital form among users located in global service areas in two modes of operation:

- “electronic mail”—with storage in the onboard memory and subsequent transmission to the user;
- real time, when the sender and recipient were in the radiovisibility area of the same satellite. Unlike Strela, the Gonets-D satellite had a magneto gravitational attitude control system, so that the long axis of the satellite was pointed at Earth with an error of 10° . The power supply system used both solar cells and a storage battery. The passive thermal control system consisted of heat pipes enabling removal of heat from the storage battery.

The main structural component uniting the blocks and assemblies of the spacecraft into a single system was the pressurized container. It served as the primary load-bearing element for connecting the spacecraft to the separation device by two brackets located diametrically on its cylindrical portion. The upper and lower frameworks for the solar cells were mounted on the upper end plate and cylindrical portion of the pressurized container, respectively. The lower framework was used as the mount for deployment of the antenna feeds and the magnetometer rods. The Gonets-D1 was launched on 2/19/96. This satellite was an updated version of the Gonets D, with an improved retransmitter to provide higher throughput. The major specifications of the Strela type (Fig. 1) and Gonets-D1 satellites (Fig. 2) are provided in Table 2.

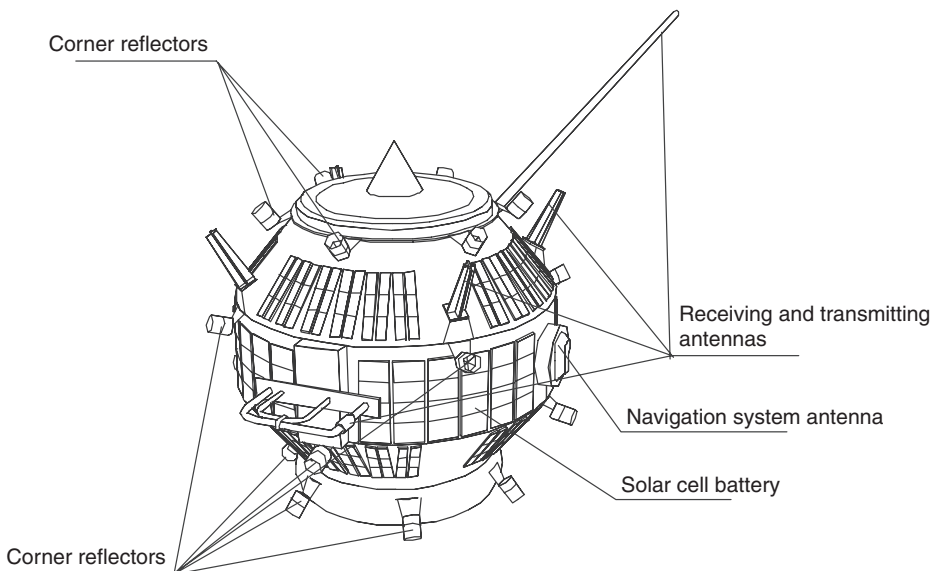


Figure 1. Strela type spacecraft.

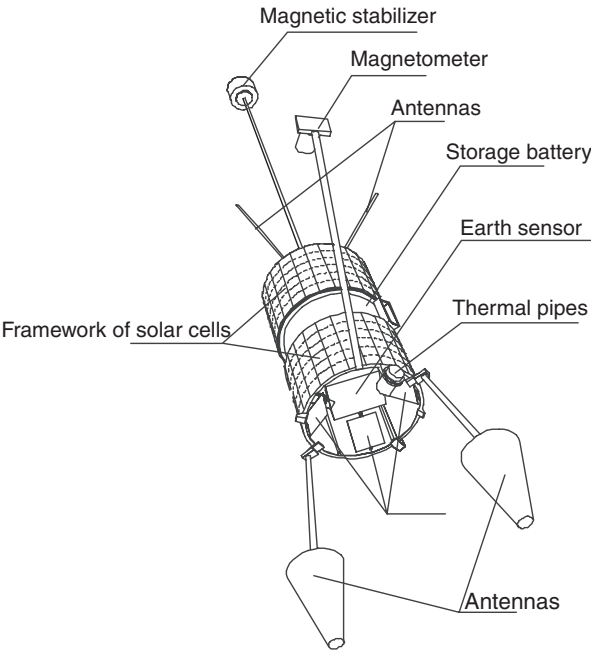


Figure 2. Gonets type spacecraft.

Satellites in Highly Elliptical Orbit

On 23 April 1965, the Molniya-1 satellite was launched, making it possible to establish communications among the most distant points in our nation. The introduction and use of this second-generation satellite enabled creation of

Table 2. Specifications of Strela and Gonets-D1 Satellites^a

	Strela	Gonets-D1
Orbit	Circular, altitude 1500 km and inclination of 85.5°	Circular, altitude 1500 km and inclination of 82.5°
Retransmitter frequency range	P	P
Position in orbit	Not oriented	Uniaxial orientation
Area of solar array, m ²	0.88	3.4
Capacity of storage battery, A-h	0.55	4
Mean power of electric power system per orbit, W	12.5	100
Type of thermal control system	Passive	Passive with heat pipes
Spacecraft mass, kg	65	225
Orbital insertion procedure	Multiple sats per launch, 5–8	Multiple sats per launch, 6

^aThe Cosmos-3M and Tsiklon launch vehicles were used to insert these satellites into orbit. The Rokot launch vehicle could also be used.

the first national satellite communication system as well as distributed television broadcasting. One feature of the highly elliptical orbit is that, during the working phase, the spacecraft has radiovisibility of the entire territory of Russia (USSR).

The Molniya-1 was the first satellite launched into a highly elliptical orbit. The orbit had a perigee altitude of 460 km, an apogee of 40,000 km, and an inclination of 62.8°. There was a retransmitter on board, which received and transmitted digital and analog data in real time, and a control and instrumentation system for transmitting telemetry information to Earth on the status of onboard systems and receiving of commands from Earth. Telemetry sensors of signals and amplitude monitored retransmitter performance during routine use.

Molniya-1 consisted of a pressurized instrument container, two antenna blocks with directional antennas, solar panels, a thermal control system heat exchanger, and a cold gas control system. Inside the pressurized container was the instrument rack and its instruments. The pressurized container maintained the interior environment (temperature, humidity, pressure, and gas composition) and was the main load-bearing component of the structure.

To transmit a continuous signal to Earth the two antenna blocks were located in a single plane on opposite sides of the spacecraft. The antenna blocks were constantly pointed at the Earth using Earth-orientation sensors.

The spacecraft got its power from an electric power system consisting of a solar panel array and a storage battery. The solar array powered the satellite equipment during the illuminated part of the orbit, and during the dark parts, the storage battery was switched on. Having the panels of the solar array constantly pointed at the Sun provided the maximum possible energy. This was achieved by adjusting of orbital parameters with a cold gas system device. A station-keeping propulsion unit that had a liquid propellant engine was used to correct orbital parameters during insertion and maintenance in orbit.

A control system was used to hook up the electric power system to the onboard equipment and to control it, as well as the spacecraft systems. The control systems consisted of a control block, contact sensors, and separation switches. The control block's function was to turn the power supply on and off, perform logic processing and distribution of individual commands from the control and instrumentation system, and also to support ground-based tests. The command segment controlling the operations of the attitude control and stabilization system, the relay station and all commands for the electric power and thermal control system were issued by the control and instrumentation system, bypassing the control block.

The contact sensors functioned to close electric circuits transmitting signals to deploy the structural elements of the antenna blocks and solar panels. The disconnect switches functioned to generate a "disconnect switch" signal at the moment the device was disconnected.

The Molniya-1 equipment consumed a great deal of electricity and thus put out a great deal of heat. A multicomponent active thermal control system maintained the thermal conditions needed for the equipment to operate. This type of thermal control system maintained the temperatures of the attachment points for the apparatus mounted on the equipment rack within the pressurized container by a flow of liquid through pipes and adjustable electric heaters.

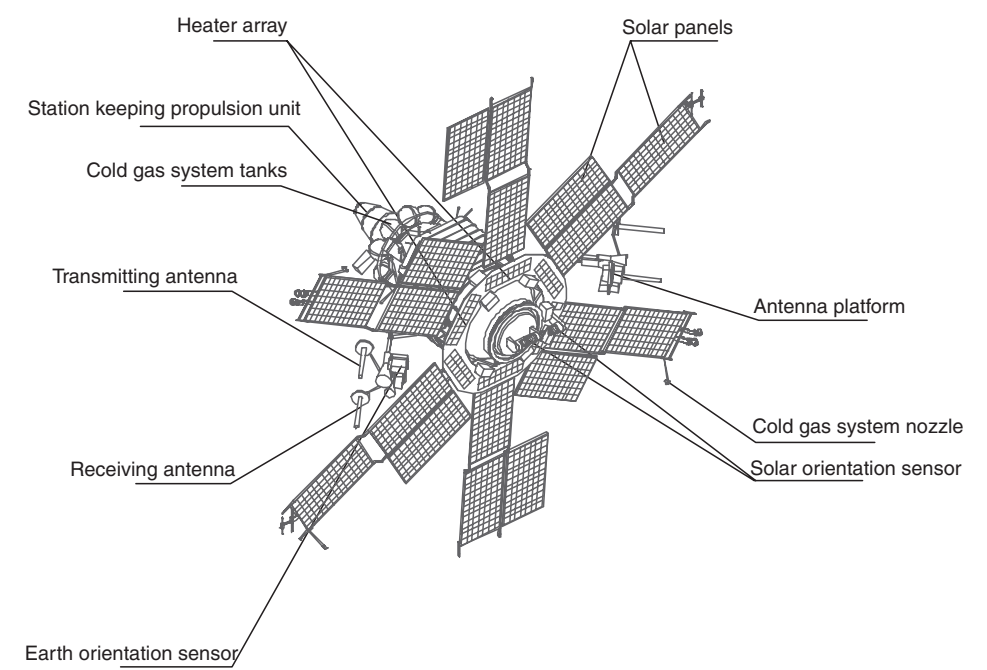


Figure 3. Molniya type spacecraft. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Using the Molniya-1 platform, the Molniya-1B was developed and launched in 1969, the Molniya-3 in 1971, and the Molniya-1 T in 1983. The structure and configuration of the spacecraft design was virtually unaltered; however, the active life of the equipment gradually increased as did channel capacity, increasing communications quality. The Molniya-3 K was launched in 2001 and has an active life span of up to 5 years. It has a new station-keeping propulsion unit, its temperature control system contains a new generation thermal regulator, new photovoltaics with improved performance have been installed on the solar panels, and the throughput capacity of the retransmitter has been improved (Fig. 3). The specifications of the Molniya type spacecraft are shown in Table 3. The Molniya launch vehicle was used to insert the spacecraft into orbit.

Table 3. Specifications of Molniya Spacecraft

Orbit	Highly elliptical at an inclination of 62.8°
Retransmitter frequency range	C, P
Position in orbit	Dual-axial orientation
Area of solar array, m ²	18
Capacity of storage battery, A-h	70
Power of electric power system, W	1400
Type of thermal control system	Active
Spacecraft mass, kg	1740
Orbital insertion procedure	Single

Spacecraft in Geostationary Orbit

The Raduga and Gorizont spacecraft were the first Russian communication satellites to operate in geostationary orbit (Fig. 4). They were designed to provide telecommunication services, including television, telephone connections, and data relay. The Gorizont, which is the most typical, was inserted into geostationary orbit by the Proton launch vehicle and the D or DM booster stage. An onboard propulsion system carried the spacecraft to the required longitude and maintained it there. Orbital inclination was not controlled. The first Gorizont spacecraft was launched on 19 December 1978, the last on 6 June 2000. During this period 33 satellites of this type were successfully launched.

The introduction of Gorizont made possible a substantial (approximately by a factor of 10) increase in the channel capacity of satellite communications in the interests of the economy and international collaboration within the Orbita and Intersputnik systems. Moreover, Moscow, a new distributive television broadcasting system, and the Volna system of communications with commercial ships at sea were developed for use on these satellites. A multi-transponder onboard retransmitter was installed on this spacecraft. The total output power of the

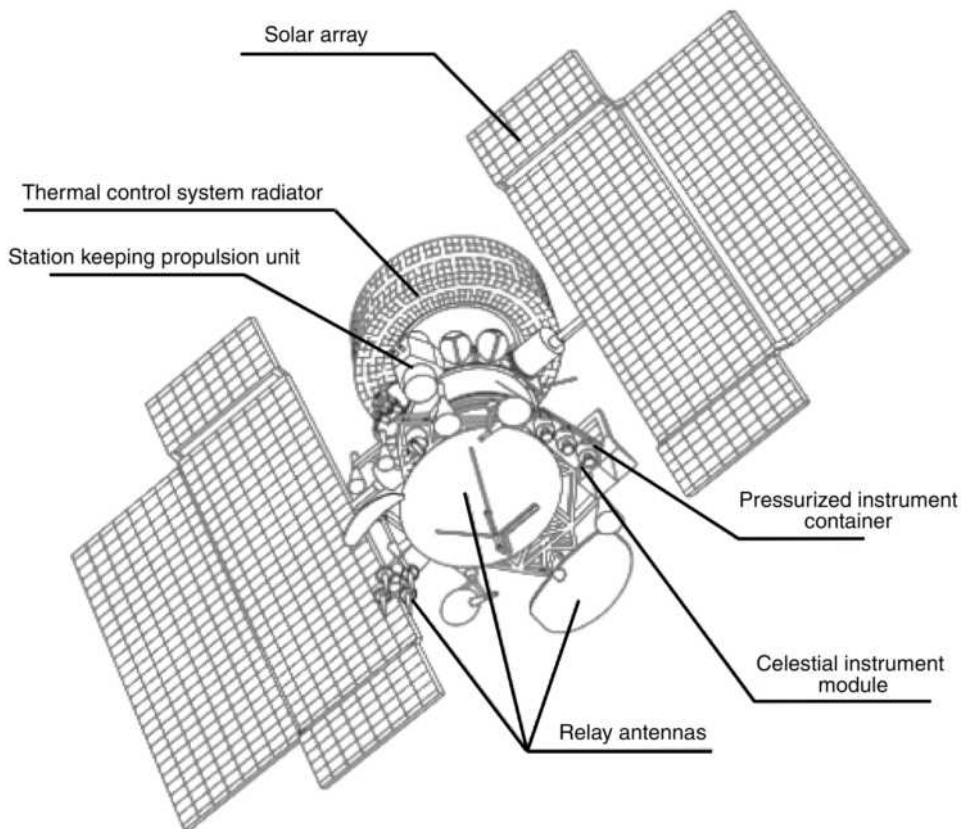


Figure 4. The Gorizont spacecraft.

onboard transponders is 195 W. The Gorizont retransmitter comprises six C-band transponders, one Ku-band transponder and one L-band transponder. The C-band transponders are multipurpose and can be used for relaying television, multichannel telephone, and radio broadcast signals, and images of newspaper matrices. One of the transponders has enhanced output capacity and is designed for transmitting television signals to a network of simplified receiving stations using a highly directional onboard antenna. The L-band transponder has a cross-link with one of the C-band transponders that allow using it for communications between a ship station in the L-band and a shore station in the C-band. Reception and transmission of radio signals is handled by a set of 11 onboard antennas.

The centimeter-band antennas provide various directional patterns and serve territories of various sizes on the Earth's surface. Regions of heavy traffic are served by highly directional antenna, and those with less traffic by global antennas. The highly directional transmitting antenna used in the Moscow system is equipped with a special device that allows it to be reaimed once. This makes it possible to serve stipulated regions on the Earth's surface in the event it became necessary to transfer the satellite from one spot in its orbit to another.

Despite the fact that Gorizont's development began in mid-1975, the basic technical designs for the onboard utility systems used in its development are still being used in current projects. The spacecraft's construction and configuration center around the pressurized instrument container. The instrument container is the main load-bearing component. It is used for mounting the structural units, instruments, and modules that comprise the satellite. The instrument container is a cylindrical pressurized module containing an instrument rack and components of the thermal control system. The interior of the container provides the most favorable climatic conditions for operating electronic instruments.

In orbit, the spacecraft is oriented so that the longitudinal axis of the instrument container, along which the onboard relay antennas are mounted, is constantly pointed at Earth, whereas the horizontal axes maintain a constant position in space. The station keeping propulsion units are located along the axis tangential to the orbit, and the axis that is binormal to the orbit corresponds to the axis of rotation of the solar panels. These panels are kept constantly pointed at the Sun by two one-stage electromechanical drives that rotate synchronously.

The spacecraft uses a triaxial attitude control and stabilization system. The sensor elements of the system include infrared Earth-orientation sensors, solar-orientation sensors, and gyroscopic angular velocity meters. The gyroscopic stabilizer and a cold gas engine act as servomechanisms. Analog devices control the system.

The orientation and stabilization system damps the movement of the satellite after it is separated from the booster stage and provides initial orientation to the Sun, initial orientation to Earth, and triaxial attitude control and stabilization of the satellite throughout its life in orbit with an error of $\pm(0.3-0.5)^\circ$.

The station-keeping propulsion units on the first Gorizont satellites utilized bipropellant liquid microthrusters that had a thrust of 45 g. The station-keeping propulsion unit had the form of two modules located so that the axes of the engine nozzles were pointed tangentially to the orbit on opposite sides. Given this configuration of the propulsion units, the onboard antennas could be pointed at

Earth without disruption, and thus it was not necessary to interrupt communication sessions when these units were on. The onboard spacecraft control system, along with a special module, controlled the propulsion unit. The station-keeping propulsion unit maintained the satellite within the stipulated longitudinal limits with an error no greater than $\pm 0.5^\circ$.

The spacecraft power supply system used a solar array with silicon photoconverters constantly pointed at the Sun, a cadmium nickel storage battery, and a voltage regulator that created optimum conditions for the common operation of the solar panels and storage batteries and stabilized the voltage in the range of (27 ± 0.81) V. The capacitance of the storage battery provided enough electric power for the onboard equipment to support uninterrupted operation of the retransmitter while the satellite was in the dark portion of the orbit. The power of the Gorizont electric power system at the end of the satellite's active life was 1280 W.

The solar array structure was redesigned to allow placing it in the limited volume of the launch vehicle nose cone during insertion into orbit. It had the form of two wings symmetrically located with respect to the instrument container. The wings were connected to electromechanical drives through articulated joints. Each wing consisted of four panels, the root, the end, and two side panels, which were connected to each other by intermediate joints. In the launch position, the wings were folded into a package and attached by special rods with a system of pins with mechanical locks that burst open after the spacecraft separated from the booster stage. The panels were rotated into operating position by spring-activated drives.

The temperature was maintained by an active, two-loop, gas-liquid thermal control system that consisted of a fan, a gas-liquid heat exchanger, a gas pipeline, a hydraulic pump, a thermoregulator, relief valves, a radiative heat exchanger, and a system control block. The required temperatures were maintained through forced gas circulation in the pressurized instrument container and forced coolant circulation through the satellite's liquid loop.

The system control block provides automatic control of the system components and, if necessary, shifts from the main components of the system to the backup components. Pressure sensors, sensors of coolant pressure differential, and check valves maintain and monitor the state of the liquid loop. The system is guaranteed to maintain the temperature in the instrument container in the range of 0 to 40°C , and the magnitude of outgoing heat flux is up to 1400 W.

The onboard control unit controls operation of the onboard spacecraft systems. The control unit consists of the equipment of the command and instrumentation system, which receives control commands from Earth, generates a signal to allow the ground control systems to locate the satellite in orbit, and transmits telemetry information on the status of onboard systems and instruments; the telemetry monitoring equipment, which provides storage and output of data from the telemetry sensors of the onboard systems to the transmitting equipment of the command/instrumentation systems; and the control block, which provides strict logic for deploying the mechanical devices and distributing electric power and control commands among the spacecraft systems and instruments.

The Ekspress and Ekspress-A satellites represent further developments of Gorizont. The development of Ekspress made maximum use of the technical experience that had been accumulated in developing the Luch-2 and Gals

spacecraft. Compared to its predecessors, *Ekspress* has significantly better technical and performance characteristics: 12 transponders for the onboard retransmitter; an improved station-keeping propulsion unit for maintaining orbital inclination, which was used for the first time on this satellite; improved precision for maintaining the satellite in orbit from 0.5° for longitude to 0.2° for longitude and latitude, and increased accuracy of spatial orientation of satellite axes in space up to 0.1 to 0.2° . With respect to design, *Ekspress* has the following distinguishing features, compared with *Gorizont*:

1. An information processing unit was added to the onboard control system for storage, maintenance, and transformation of control information to support satellite control, function, and onboard software. This software consists of a package of programs, which, in accordance with algorithms, receives information from the sensor systems, performs calculations, distributes the results, and controls the system or servomechanisms, or prepares information for other programs. The main equipment is the onboard digital computer. These control information processing and computing systems were first used on the *Potok* and *Luch* spacecraft.
2. The attitude control system has improved attitude control instruments, including a system of instruments that provides orientation to the North Star. Logic control over the system is the function of the control information processing and computing complex, which is part of the onboard control system. These measures increase the accuracy of the orientation of the satellite's axes in space to 0.1 – 0.2° .
3. New, more effective types of station-keeping and attitude control engine units were adopted. As a result of the new task of correcting orbital inclination and the associated need to produce significant thrust, with limited mass resources available, *Ekspress* uses a station-keeping propulsion unit, including stationary plasma engines whose specific impulse characteristics exceed the analogous performance of other types of engines by an order of magnitude ($I_{sp} = 1500$ s). This has made it possible to reduce the propulsion unit mass to a minimum by using only a small quantity of working fluid—xenon—and a small number of tanks to store it. Propulsion units of this type were first used on the *Potok* satellite. For orientation, because of the relative frequency with which the engines have to be turned on, a propulsion unit using thermocatalytic engines of relatively high specific impulse performance ($I_{sp} = 170$ s) and low power consumption is used.
4. The electric power system uses a storage battery with nickel hydrogen storage cells with high specific energy and efficiency characteristics compared to the cadmium nickel storage cells used on *Gorizont*.
5. *Ekspress* is similar in design to *Gorizont*, except for the location of the transmitters of the onboard retransmitter, which were removed from the pressurized instrument container and placed on the exterior framework where all of the onboard retransmission antennas are located.

A modification of this spacecraft—*Ekspress-A*—has 17 transponders and a 7-year active lifetime in orbit.

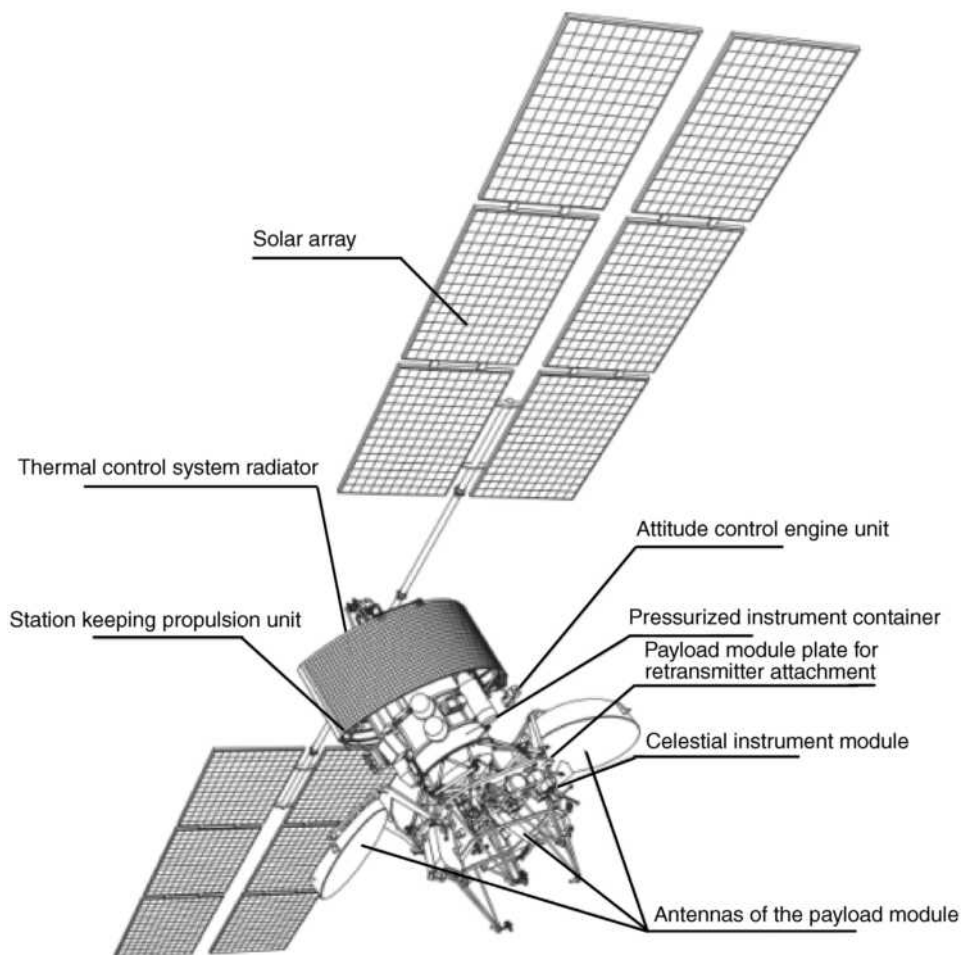


Figure 5. SESAT spacecraft.

Table 4. Specifications of Geostationary Spacecraft

	Gorizont	Ekspress-A	SESAT
Throughput, number of transponders (output power, W)	8 (195)	17 (510)	18 (1512)
Orientation precision, degrees	0.3–0.5	0.1–0.2	0.1
Longitude hold, degrees	0.5	0.2	0.1
Latitude hold, degrees	–	0.2	0.1
Power capacity, W	1280	2900	5300
Total impulse of station keeping propulsion, kN.s	118	1200	1600
Cold generation by the thermal control system, W	1400	2150	2550
Days of autonomous operation	–	30	–
Active lifetime in orbit, yr	3	7	10
Mass, kg	2200	2600	2600

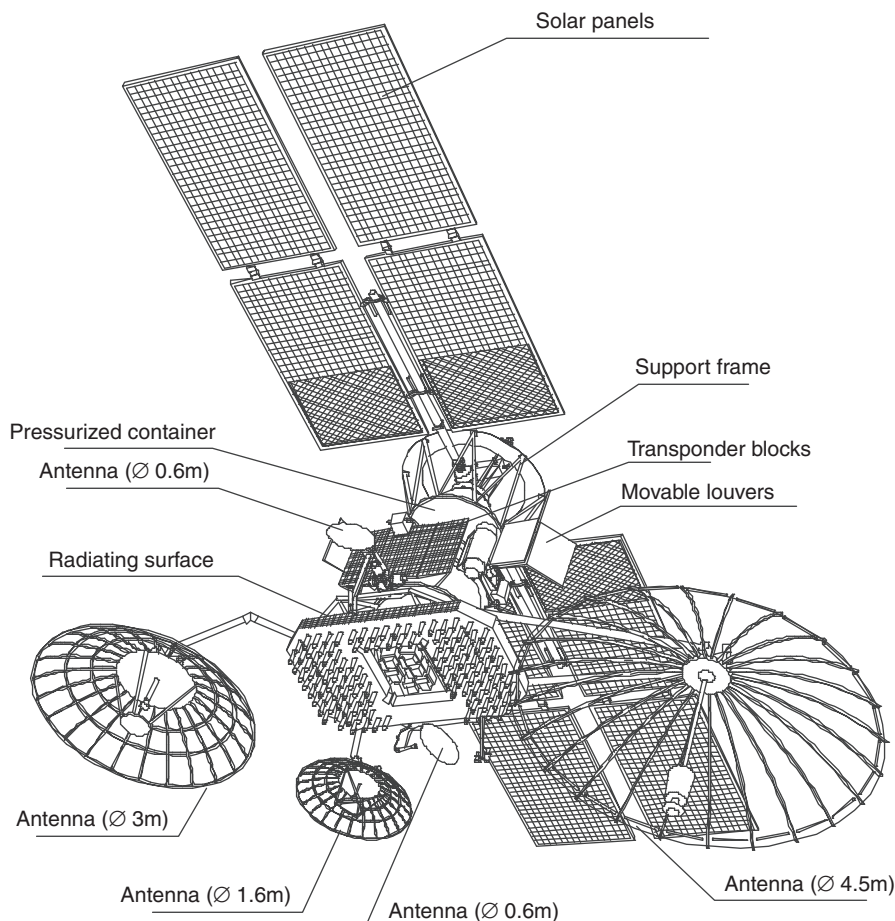


Figure 6. Luch spacecraft.

A further step in the development of telecommunication satellites in this series is the SESAT spacecraft, which has a 10-year active orbital life time and 18 transponders in the Ku-range (Fig. 5). The Proton launch vehicle is used to insert all these types of spacecraft into orbit. The major specifications of the geostationary spacecraft developed by NPO PM are provided in Table 4.

Table 5. Specifications for Luch Spacecraft

Number of transponders	7
Retransmitter frequency band	Ku, P
Orientation precision, degrees	0.1–0.2
Longitude hold, degrees	±0.5
Deviation in latitude, degrees	±3.25
Electric power, W	1800
Spacecraft mass, kg	2400
Active lifetime, years	5

The geostationary satellites Luch are also telecommunication satellites developed by Russia. They are part of the satellite system of command and control of piloted and unpiloted spacecraft. Their development used technical designs validated on other spacecraft developed by NPO PM, such as Gorizont and Potok (Fig. 6). Their distinguishing feature is an antenna system with a high degree of transformability in the launch position, which is comprised of various different types of antennas. The electromechanical pointing system, which uses BCVM algorithms, makes it possible to track the user's low-orbital spacecraft. The specifications of this spacecraft are provided in Table 5.

A.G. KOZLOV

V.A. BARTENEV

Scientific-Production Association
of Applied Mechanics
Russia

COMMUNICATIONS SATELLITES, TECHNOLOGY OF

History

Early Theories. In 1895, a Russian scientist, Konstantin Tsiolkovsky, gave the world its first vision of a stationary satellite. He observed that an object orbiting Earth at 22,300 miles up would match the angular rotation of Earth and thus provide a seemingly stationary "star" overhead.

Fifty years later, the English science fiction writer and Royal Air Force electronics officer Arthur C. Clarke expanded on this vision. In 1945, he indicated that such an object orbiting Earth at 22,300 miles up must also have its orbit in the equatorial plane to be considered stationary. Clarke called the object a "satellite" and further noted that providing a satellite in this orbit with a communications repeater could produce a very valuable communications capability. Clarke named such satellites "geostationary communication satellites." Clarke postulated that three geostationary satellites spaced 120° apart would provide full global communications coverage for telephone and television service to the world's population (1,2).

Clarke's concept was ahead of the technologies needed to make it a reality. In 1945, the science of rocketry was still crude, and radio communications depended on large, unreliable vacuum tubes. However, rocketry and electronics technologies both began to advance rapidly, particularly after the development of the transistor in 1947. The first satellite launches, into a low Earth orbit (LEO), occurred during the Cold War following World War II. The United States and Soviet Union were intent on achieving military and technological superiority over each other. Satellites were seen as potentially valuable Cold War tools for reconnaissance and for maintaining reliable communications among far-flung troops.

The Earliest Satellites. On 4 October 1957, a Soviet satellite named “Prosteyshiy Sputnik” (“Simplest Satellite”) was launched into a low Earth orbit and began transmitting telemetric information for what would turn out to be 21 days. Sputnik weighed 184 pounds and was a sphere two feet in diameter. This dramatic Soviet achievement shocked the world, and particularly the U.S. public, which perceived it as a threat to national security. The event was likened to the surprise attack on Pearl Harbor, and it sparked a flurry of U.S. government activity in rocketry and satellites (3).

The rocket that lifted Sputnik into orbit was four stories high. Such heavy-lift launchers used technology similar to that of intercontinental ballistic missiles (ICBMs), so it was perceived that the launch vehicle was at least as important as the actual satellite. In the United States, the public feared that the Soviet Union’s ability to launch satellites translated into the capability to launch ballistic missiles that could carry nuclear weapons from Europe to America. The “Space Race” had begun (4,5). The Sputnik launch led directly to the creation of the U.S. National Aeronautics and Space Administration (NASA) (6). In October 1958, the National Aeronautics and Space Act launched NASA out of the old National Advisory Committee for Aeronautics (NACA), which had existed since 1915. NASA became responsible for civilian space science and aeronautical research, whereas the U.S. Department of Defense continued to carry out defense work.

The launch of Sputnik II on 3 November 1957, was a second blow to America’s perception of itself as technologically superior to the Soviet Union. A third blow came on 6 December 1957, when the first U.S. satellite, Vanguard, exploded on launch. But on 31 January 1958, the United States launched Explorer I, which successfully transmitted telemetric data for 5 months. Shortly thereafter, in December 1958, America launched Score, the first satellite to transmit a radio broadcast, President Eisenhower’s Christmas message. In 1960, two more satellites followed, Echo I and Courier. On 10 July 1962, NASA launched the first non-government-built satellite, AT&T’s Telstar 1, a LEO spacecraft for relaying transatlantic television and data. In that same year, Telstar was used for the first transoceanic television ever transmitted, a live broadcast that commemorated the first anniversary of the death of U.N. Secretary General Dag Hammarskjöld by a link-up of simultaneous ceremonies held at the United Nations in New York and Paris and Hammarskjöld’s tomb in Sweden (7). Relay I, launched by NASA in December 1962, was built by RCA, also for transoceanic communications. Meanwhile, in May 1958, Sputnik III was placed into orbit by the Soviet Union. This was a large satellite for the time and demonstrated that the Soviets were ahead of the United States in heavy-lift rocketry.

Geosynchronous Satellites. All of these early satellites were nongeostationary and nongeosynchronous [a geosynchronous (GEO) satellite orbits at 22,300 miles, but is not necessarily geostationary, that is, limited to the equatorial plane—see discussion later]. They were launched into low-altitude Earth orbits because the rockets of the day could not propel the satellites into an orbit 22,300 miles up.

In the early 1960s, when U.S. rockets could at last boost a satellite into geosynchronous orbit, one of the most important questions centered on which was the best orbit to use for a communications satellite, low Earth orbit or geosynchronous. Low-altitude systems had the advantages of lower launch costs,

heavier payloads, and relatively short radio-frequency propagative times. The main disadvantages were that many orbiting satellites were required to achieve continuous global communications, and these needed continuous tracking. Geosynchronous satellites, in contrast, had two key advantages. Only three satellites were needed for global coverage, and only minimal tracking was required. Their one primary disadvantage was relatively long radio-frequency propagative times. No one knew how the one-quarter-of-a-second transmission delay would affect the feasibility of using this orbit for telephony (see extended discussion later).

By 1959, a small team of scientists led by Dr. Harold Rosen from the Hughes Aircraft Company (now Hughes Electronics Corporation) was moving ahead, determined to create a geosynchronous communications satellite. By 1960, they had built a satellite prototype (8,9). Meanwhile, at AT&T's Bell Laboratories, similar work was taking place under the direction of Dr. John R. Pierce and with at least as much zeal. Arthur Clarke would later name John Pierce and Harold Rosen the "fathers of communications satellites." Pierce's team demonstrated the first active communications repeater, but Rosen and his team at Hughes are credited with making it possible, technically and economically, to have a continuous communications capability by satellite earlier than anyone thought feasible (10).

In August 1961, NASA contracted with the Hughes Aircraft Company for the first geosynchronous communications satellite, called Syncom (Fig. 1). Though the first Syncom launch failed in February 1963, a second attempt, the launch of Syncom II in July of the same year, succeeded. Syncom I had just one voice channel and was designed to weigh 86 pounds at the beginning of its life. Though it never attained orbit, it paved the way for Syncom II and Syncom III, satellites that, by August 1964, proved the feasibility and cost-efficiency of domestic and international satellite communications. The Syncom series also validated the concept of geosynchronous satellites, and by 1964, government and other users turned away from LEO satellites for voice, data, and video communications (11).

COMSAT, INTELSAT, and INMARSAT. The 1962 U.S. Communications Satellite Act had a profound impact on international satellite communications. It provided for the establishment of the Communications Satellite Corporation (COMSAT) (12), a privately financed and managed organization that had a minority of U.S. government representatives on its board of directors. AT&T emerged as COMSAT's largest shareholder. COMSAT was created (1) to govern the operation of communications satellites and ground facilities used to transmit to and from the United States and (2) to develop and manage a new international communications satellite organization, which, in 1964, emerged as the International Telecommunication Satellite Organization, or INTELSAT. COMSAT was responsible for the procurement, testing, and launch acquisition of all INTELSAT satellites, and it owned 61% of the organization (13,14). The global, commercial INTELSAT cooperative was formed on the initiative of U.S. President John F. Kennedy, who, at the time COMSAT was created, said, "I invite all nations to participate in a communications satellite system in the interest of world peace and closer brotherhood among peoples of the world." INTELSAT was the first organization to provide global satellite coverage and connectivity. In its role as a commercial cooperative and wholesaler of satellite communications capacity, INTELSAT provides service through its signatories in member



Figure 1. Hughes engineers Dr. Harold Rosen (right) and Thomas Hudspeth hold a prototype of the geosynchronous Syncom satellite atop the Eiffel Tower during the 1962 Paris Air Show (courtesy Hughes Electronics Corp.).

countries. Currently, COMSAT is the only U.S. signatory, though pending deregulation will encourage others (15).

COMSAT's initial capitalization of \$200 million was considered sufficient to build a system of dozens of medium Earth orbit (MEO) satellites. These orbit the Earth at about 3000 to 7000 miles up; fewer MEOs are needed for global coverage than LEOs. In 1964, when COMSAT was in the process of contracting for its first satellite, two Telstars, two Relays, and two Syncoms had operated successfully in space. For a variety of reasons, including cost, COMSAT ultimately rejected a joint AT&T/RCA bid for a MEO system that incorporated the best of Telstar and Relay. Instead, COMSAT chose the geosynchronous satellite offered by Hughes, based on Syncom technology. Procured by COMSAT but transferred to the newly formed INTELSAT, the Early Bird satellite (also called "INTELSAT I") was launched on 6 April, 1965, as the first commercial communications satellite. Built by Hughes, it was launched from Cape Canaveral on a Delta rocket and weighed 85 pounds—still about all that could be lifted to a geosynchronous orbit at that time by American technology.

Early Bird was designed to test the feasibility of synchronous orbits for commercial communications satellites and was a resounding success. The

satellite provided 240 transatlantic telephone circuits capable of carrying that many calls simultaneously. This greatly increased telephone capacity across the Atlantic. Early Bird could provide almost 10 times the capacity of a submarine telephone cable for almost one-tenth the price and thereby helped prove the cost-efficiency of communications satellites. Moreover, the public readily accepted the transmission delay. Early Bird was also designed for transmitting television. Though built for just 18 months of life, it could still transmit live pictures of the Apollo Moon landing in 1969. By 1967, three of these first-generation INTELSAT satellites were operating over the Atlantic and Pacific oceans, providing ubiquitous global telephone and television communications for the first time (16).

In February 1976, COMSAT launched a new kind of communications satellite, Marisat, to provide mobile communications services to the U.S. Navy and other maritime and aeronautical customers. Subsequently, in 1979, COMSAT transferred Marisat to the newly formed International Maritime Satellite Organization, INMARSAT. Sponsored by the United Nations, INMARSAT has an intergovernmental structure similar to INTELSAT's. Each signatory is required to provide an interface with land-based telecommunications networks and is assigned an investment share based on its actual use of the system. Today, COMSAT manages access to the global satellite fleets of both INTELSAT and INMARSAT on behalf of U.S. and foreign telecommunications operators that want to initiate or terminate their satellite transmissions in the United States. COMSAT currently enjoys an exclusive relationship with both INTELSAT and INMARSAT in providing this service.

As of June 1999, INTELSAT owned and operated 24 satellites and had a membership of 143 countries and signatories. The highly successful organization, expected to be privatized soon, provides voice, data, video, and Internet services on a nondiscriminatory basis to more than 200 countries and territories. Since its formation, INMARSAT has greatly expanded its services. Today, the organization provides satellite global mobile communications services on land, sea, and in the air. For maritime users, INMARSAT supports phone, telex, fax, electronic mail, and data transmission. Aeronautical applications include flight-deck voice and data, automatic passenger telephone, fax, and data communications. INMARSAT's land-based customers have access to in-vehicle and transportable phone, fax, and two-way data communications, position reporting, electronic mail, and fleet management for land transport. INMARSAT is also available for disaster and emergency communications, as well as news reporting, where alternative communications links are difficult to access or nonexistent (17). Initially, INMARSAT leased transponders on other organizations' satellites, but in October 1990, it launched the first of its own satellites, INMARSAT II-F. As of June 1999, INMARSAT had 85 member nations and a global fleet of four geosynchronous satellites.

The arrangement, mentioned earlier, whereby COMSAT enjoys monopoly status as an access point for U.S. domestic and international communications organizations, along with other "privileges and immunities" conferred on it by the U.S. government, is currently being challenged. In fact, there is a worldwide effort underway to push COMSAT, INTELSAT, and INMARSAT to serve their customers on a fully commercial basis, free of all government ties and protective

legislation and regulations. Both INTELSAT and INMARSAT have embarked on this path with the New Skies and ICO satellite systems, respectively. Today, COMSAT is partly owned by Lockheed Martin, a U.S. aerospace corporation and satellite manufacturer. As of June 1999, Lockheed Martin was seeking to buy all of COMSAT, for which it needs regulatory approval.

It is inevitable that INTELSAT will evolve to privatized organizations. These will remain one of history's best examples of successful international cooperation.

A Worldwide Industry. The use of communications satellites proceeded apace in the United States and other countries. In a number of cases, satellites offered cost savings and improved signal quality, compared with terrestrial and submarine cables. Importantly, satellite transmission was cost-insensitive to distance, as opposed to the way terrestrial transmission facilities were priced. Although the initial commercial launch vehicles and satellites were American, other countries had been involved in commercial and government satellite communications from the beginning. For example, as early as 1961, the German Post Office announced plans to construct a satellite-receiving ground station that could handle up to 600 phone calls simultaneously. The ground station interfaced with Telstar and Relay-type satellites. By the time Early Bird was launched, satellite Earth stations already existed in the United Kingdom, France, Germany, Italy, Brazil, and Japan.

In 1969, the Canadian government created Telesat Canada, a corporation charged with building and operating the world's first national satellite network. The first satellite, Anik A (Anik means "little brother" in the Inuit language), was launched in November 1972, and a three-satellite system was completed by May 1975. Anik was notable for introducing new spacecraft technologies, such as the shaped beam. Syncom, Early Bird, and Intelsat II used global beam technology, meaning that the satellites broadcast their signals across the entire third of the world they faced. Anik's shaped beam permitted a pattern, or footprint, that covered only Canada (18). Following the success of Canada's Anik system, both Europe and the United States developed domestic satellite systems. In 1974, Europe's first telecommunications satellite, Symphonie, was launched. In the United States, Western Union contracted with Hughes for three Anik-type satellites to serve the U.S. market. The first of these, Westar I, was launched in April 1974. In the following year, RCA launched its network system, and in 1975, the first live commercial television program was transmitted in the United States. Also in 1975, Home Box Office became the first national cable TV network to be delivered by satellite—and thus began a whole new entertainment industry. By the end of 1976, there were 120 satellite transponders available in the United States; each could provide 1500 telephone circuits or one TV channel (19).

Indonesia, the world's largest archipelago of more than 13,000 islands, launched its first Palapa satellite in 1976 and followed with another in 1977. Besides supporting telecommunications, Palapa-A1 and Palapa-A2 were intended for distance learning, including teaching the national language (20). By the mid-1980s, many nations had national satellite systems. These included Australia (OPTUS), Mexico (MORELOS), and Brazil (BRASILSAT). In 1999, this tally had grown to include most of the world's developed and developing countries. In addition, there are numerous private regional satellite systems, such as

Eutelsat (Europe); SES Astra (Europe); Europe*Star; Nahuel (Central and South America); Galaxy Latin America; Thuraya (Middle East and parts of Asia, Africa, and Europe); ACeS (Asia); and AsiaSat, which delivers service to China, Thailand, Malaysia, Pakistan, Hong Kong, Burma, and Mongolia (21).

Today, INTELSAT continues to own and operate the world's largest global satellite communications system (Fig. 2). However, there is now strong competition from such international entities as PanAmSat, which, it is projected will overtake INTELSAT in the near future. Loral, GE Americom, and Intersputnik also operate global fleets. Despite the early failure of Iridium (Motorola), Globalstar (Loral/Qualcomm) and ICO (a private INMARSAT spinoff) promise to provide worldwide satellite mobile services. Spaceway (Hughes), Teledesic (Boeing, Motorola, Matra Marconi Space), and Astrolink (Lockheed Martin, TRW, Telespazio) are among several companies that anticipate operating global satellite systems for high-speed data exchange, Internet applications, and interactive multimedia.

Satellite manufacturers and launch vehicle providers are found throughout the world, too. Satellite manufacturers in the United States, include Hughes, Loral, Lockheed Martin, TRW, Orbital Sciences, and Motorola. They are joined

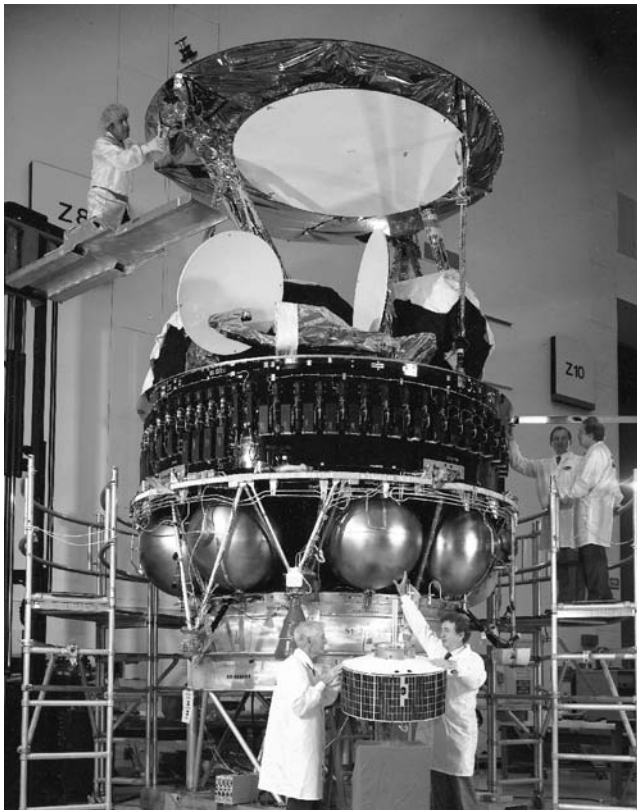


Figure 2. The tiny Syncom satellite (foreground) would fit into one of the INTELSAT VI fuel tanks to which Dr. Harold Rosen is pointing (courtesy Hughes Electronics Corp.). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

by non-U.S. competitors such as Daimler-Chrysler Aerospace (Germany), Alenia Spazio (Italy), Aerospatiale (France), Matra Marconi (France) with British Aerospace (UK), Thomson CSF (France), and Alcatel (France). In the United States, launch vehicles are provided by Boeing (Delta and Sea Launch) and ILS-Lockheed Martin (Atlas). Other prominent launch vehicles include Ariane (France), Long March (China), ILS Proton (Russia), H-IIA (Japan), and Soyuz (Russia). It is likely that there will be consolidation of the many satellite and launch vehicle manufacturers and service providers in the future.

Today, communication satellites are a basic element of the world's telecommunications infrastructure and an indispensable tool of the global marketplace. More than 325 commercial communications satellites—mostly GEOs—now orbit Earth. New applications for these satellites and those in LEO or MEO orbits are constantly evolving, based on new technological developments and increasing demand for space-based communications. Among the newest applications are satellite mobile radio, high-speed broadband data exchange, interactive multimedia, handheld global mobile telephony, and direct-to-home TV featuring high-definition pictures—HDTV (see fuller discussion in the article on commercial applications). These business and consumer services are provided through GEO as well as LEO and MEO satellites.

The early satellites had one television channel, weighed less than 100 pounds, and could require a 30-meter diameter antenna for reception. Today, a communication satellite can transmit more than 200 digital television channels, weigh 10,000 pounds, and deliver signals to one-half-meter Earth stations or palm-sized receiver/transmitter units.

Satellite Technology

Communications Satellites—Yesterday and Today. In a few decades, communications satellites have become an indispensable part of the world's telecommunications infrastructure. They serve as repeaters in the sky and relay radio transmissions from/to terminals on Earth and in space, much as terrestrial microwave repeaters receive, amplify, and retransmit signals on Earth. During the years since Early Bird, communications satellites have grown in many aspects—size, weight, lifetime, power, and capacity. One way of illustrating this growth is by comparing Early Bird to a satellite designed for the early twenty-first century. The drum-shaped Early Bird was 71 cm (28 in) in diameter and only 58 cm (23 in) high. Its mass in orbit was 34.5 kg (76 lb). By contrast, a late-model, high-powered satellite called the HS 702 towers over Early Bird. In its stowed configuration, this Hughes-built satellite measures 2 m (6 ft 7 in) \times 3.15 m (10 ft 6 in), and rises to a height of 3.6 m (11 ft 11 in), not including its antennae. Deployed in space, it spans 40.9 m (134.5 ft) in length (Fig. 3). Its mass at launch is as much as 5200 kg (11,464 lb), depending on the type of fuel. Its dry weight (which excludes fuel) is as much as 3450 kg (7590 lb), two orders of magnitude greater than that of Early Bird. Early Bird was designed for 18 months of life, the HS 702 is designed for missions 15 years long. It can generate as much as 15 kW of direct current (dc) power, three orders of magnitude more than Early Bird. Its transponder capacity is nearly 100 times that of Early Bird.

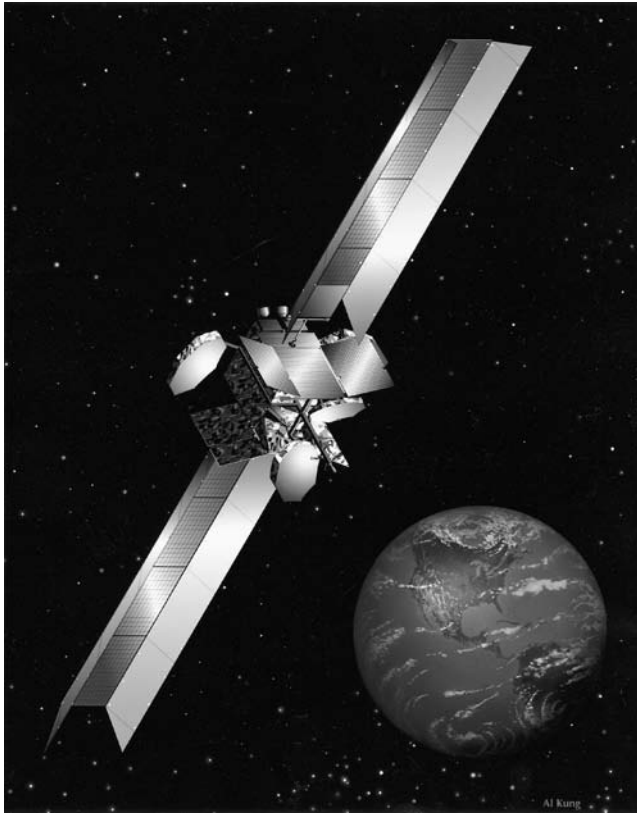


Figure 3. Artist's concept of the Hughes-built HS 702 satellite deployed in space shows how the long solar panels collect light that is used to generate 15 kW of dc power (courtesy Hughes Electronics Corp.). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Orbits. Communications satellites travel in space along a gravity-induced path, or orbit, much as natural satellites orbit celestial bodies throughout the solar system (22–24). For the first 30 years of commercial service, communications satellite designers relied almost exclusively on geosynchronous or geostationary orbits. To observers on the Earth, a geostationary satellite appears to be “fixed” in space, although it is actually traveling at a speed of 11,062 km/h (6875 mph). From a position in geostationary orbit, approximately 35,880 km (22,300 mi) above the equator, a single satellite can illuminate a large oval-shaped area on Earth that is roughly one-third of the globe’s surface. Because the GEO satellite is “fixed” in space, ground stations within the illuminated area do not have to rotate their antennae to track the satellite, as they historically had to do for satellites in non-GEO orbits. Eliminating this bit of complexity was significant, especially in the early years when the antennae of ground terminals had to be large—of the order of 30 m (98 ft) in diameter—to receive weak signals from low-powered satellites. Large antennae were necessary to compensate for the loss of energy transmitted across the lengthy geosynchronous distances because energy received by a ground antenna is inversely proportional to the square of

the distance from the transmitting source. As the power transmitted from GEO satellites has increased over the years, satellite system operators have been able to reduce the size of ground antennae significantly.

At times, satellites can be operated usefully from nonstationary geosynchronous orbits. For example, when onboard fuel is nearing exhaustion, a GEO satellite can be allowed to drift north and south of the equator while still maintaining its longitudinal position. Under these conditions, its orbit develops a small inclination with respect to the equator that grows about 1° a year due to the pull of solar and lunar gravity. For some commercial applications, like TV distribution to cable operators, departure from the satellite's "fixed" north-south position is acceptable. In this event, the satellite would appear to a ground observer to trace out a figure eight pattern in the sky; the midpoint of the eight is centered on the equator.

Until the advent of personal mobile satellite communications late in the 1990s, all commercial and most military communications satellites operated in geosynchronous orbits (25). The GEO orbit is also favored for government weather, communications, and defense early warning satellites. However, a few of the first experimental military communications satellites and several precursors of today's commercial communications satellites were placed in much lower altitude orbits. The first version of the U.S. Defense Department's Strategic Communication Satellite System (DSCS 1) operated from an orbit that combined features of medium-altitude and geosynchronous orbits. A variety of other orbital geometries that range across many combinations of altitudes and inclinations, including widely employed 90° polar orbits, have been popular for weather, navigational, Earth resources, and surveillance satellites. The satellites of the Global Positioning System (GPS), for example, operate in six planes defined by 12-hour circular orbits—each plane separated from its neighbor by 55° —and are located 20,178 km (12,541 mi) above Earth (26).

In recent years, a number of non-GEO orbits have gained currency for such applications as personal mobile satellite communications (27,28). These include LEO and MEO orbits. The altitude of LEO orbits usually ranges from about 500 km (311 mi) to 1500 km (932 mi) above Earth, whereas MEO orbits are roughly 5000 km (3,107 mi) to 12,000 km (7457 mi). Orbital configurations at these lower altitudes are carefully chosen to avoid exposing a satellite's electronic components to life-limiting radiation in the Van Allen belts, unless the operator is willing to bear the cost and weight of radiation shielding. The belts are between about 2000 km (1243 mi) and 8000 km (4971 mi), and also above MEO, but well below GEO altitude. Still another type of orbit, the high Earth orbit (HEO), a form of the highly elliptical, 12-hour Molniya orbit extensively employed by reconnaissance satellites during the Cold War, has been adapted for broadcasting CD radio from space.

The historical attraction of GEO orbits is that they can provide service with few satellites and operate through fixed antennae on the ground. Moreover, they provide ubiquitous service for users throughout their broad coverage area. A satellite operator can offer service across nearly a third of the globe using one satellite—and can serve the entire Earth, excluding sparsely populated polar regions, with three. As the altitude of an in-orbit satellite system decreases, more satellites are needed for continuous global coverage because of the shrinking

area they can illuminate on Earth at any given time. However, the satellites themselves can be smaller and lighter, and therefore less costly than modern GEO satellites (29). The typical LEO satellite is shorter lived because of the effects of atmospheric drag and radiation at low altitude. Hence, it needs to be replaced more frequently than its GEO counterpart, adding to system cost. The cost of launching LEO satellites, however, is lower because they require less energy to boost them into low-altitude orbits. But the lower the altitude, the greater the need for extensive ground networks or complicated signal handoffs from one satellite to another, as the satellites pass within view.

This is so because lower altitude systems remain in view of the mobile satellite system user for very brief intervals, perhaps only minutes. By contrast, higher altitude satellites pass much more slowly through the viewer's field of vision. Consequently, at lower altitudes, a typical call will have to be handed over from one satellite to another more frequently than at higher altitudes. The greater the frequency of handovers, the greater the chance of a phone disconnect. Often, low-altitude satellites will appear close to the observer's horizon, increasing the likelihood that an obstruction such as a building will interrupt the call. This in turn means that the higher the satellite's elevation angle (i.e., the angle from the observer to the satellite) at any given time, the less likely that the call will be lost due to surrounding obstructions. Lower altitude satellite systems add to the cost and complexity of the ground infrastructure.

Communications Links, Frequencies, and Bandwidths. The communications link between satellite and ground station can be represented by a power balance equation expressed logarithmically in decibels (dB). This relationship takes into account the predictable factors in the communications loop. It indicates that the power received at the receiver is equal to a summation of all of the gains and losses in the link. Thus, power received equals the transmitted power, minus the waveguide losses, plus the gain of the transmitting antenna, minus the propagative path losses, plus the gain of the receiving antenna, minus the receiver waveguide losses. The free-space path loss is substantial and is fixed for a given frequency and distance. At GEO altitude, it amounts to 200 dB at the 12-GHz transmission range of a direct broadcast satellite and 196 dB at the 4-GHz band of a video distribution satellite (30). The measure of the difference between the power actually received and the threshold power required for reception is called the link margin. System designers try to provide sufficient link margin to ensure successful communications.

Frequencies allocated for use or potential use by satellites extend through a number of bands, or range of frequencies (31). Within those bands, separate portions are allotted for communications uplinks to and downlinks from the satellites to keep the two separated from one another. The satellite's transponder translates signals from the uplink frequencies into downlink frequencies. The bands are identified by letter designations, a practice derived from the World War II lettering scheme for military electronic equipment. Thus, satellite frequencies extend from longer wavelengths at L band up to very short wavelengths at Q band (32). Satellite communications service for mobile users, such as those aboard ships, barges, and oil rigs, are in the L band (1–2 GHz), where they have replaced unreliable short-wave radio. Other mobile services, including personal mobile communications, are in the S band (2–4 GHz) as well as the L band. Fixed

satellite service (telephony, data, and facsimile communications through Earth stations at fixed locations) is in the C (4–8 GHz), Ku (12.5–18 GHz), K (18–26.5 GHz), and Ka (26.5–40 GHz) bands. Broadcast satellite service (direct-to-user TV) is in the Ku and K bands. Military communications satellites use the X band (8–12.5 GHz) for fixed satellite service and Ka and ultra-high-frequency (300 MHz to 3 GHz) bands. Other bands in the high millimeter range, V (40–50 GHz) and Q (above 50 GHz) have promising applications for transmitting large quantities of information to small antennas.

Frequencies in the Ku, L, and C bands are regarded as prime real estate for communications satellites. C-band frequencies are most commonly employed by commercial communications satellites. Among available frequencies, the C band is least affected by man-made noise and atmospheric attenuation. But where broadband and high capacity are needed, the higher frequencies located in the Ku and Ka bands become more desirable. Portions of the Ku band, in particular, are widely used for business communications through small Earth terminals and for high-power TV broadcasting. The shorter frequencies (longer wavelengths) below the L band are subject to ionospheric disturbances that cause fading and other random signal disruption. This explains why L-band mobile satellite communication service quickly supplanted short-wave communications for ship-to-ship and ship-to-shore applications in the 1970s.

Bandwidth, the capacity of a satellite communications link, is the spread of usable (i.e., as allocated by regulation) frequencies available for transmitting intelligible information. Available bandwidth is divided among the transponders contained in the satellite's communications payload. Transponders (or repeaters) are electronic devices that receive radio signals, amplify them to increase their strength, and transmit them on command at different specific frequencies (33). At the heart of the transponder are either traveling-wave tube amplifiers (TWTAs) for high-power, higher frequency applications or solid-state power amplifiers (SSPAs) for lower frequency, usually lower power applications. Communications satellites that orbited in the late 1990s typically carried 48 transponders. The newest satellites entering service at the turn of the century offer as many as 94 transponders, plus backups. By the late nineties, the satellite fleets of major international satellite service providers such as INTELSAT and PanAmSat each provided service via about 800 orbital transponders.

How bandwidth is divided among transponders reflects a balance of factors, including the satellite's power resources, the number of TWTAs that can be carried, and the desired power per channel, or power per transponder. The following is an illustration. A satellite that can generate 8000 W of dc power might supply 7000 W to its communications payload, mainly for the TWTAs. If the TWTAs' portion of that figure—say, 6800 W—is converted to radio-frequency (RF) power at 60% conversion efficiency, there will be 4080 W of RF power available from the TWTAs. If satellite designers want 100 W of power per channel, there is sufficient power for 40 channels. For 200 W per channel, 20 channels can be created. If the choice is 20 channels—10 channels for each of two polarizations—the total available bandwidth will be divided by 10. If the bandwidth (capacity) available within the allotted “usable frequency” range is 250 MHz, there will be 25 MHz available per channel. Some of that bandwidth, perhaps 10%, has to be reserved for guardbands that separate the channels.

Consequently, the usable bandwidth will be 0.9 times 25, or 22.5 MHz per channel.

Onboard Antennas and Spot Beams. The area on Earth illuminated by communications satellites, or coverage area, is commonly referred to as the satellite's footprint. To improve efficiency, satellite designers usually configure antenna patterns to illuminate specific high-revenue-producing areas on the ground, or a particular country or group of countries for a domestic or regional satellite system. Traditionally, shaping beams to match the contours of specific areas on Earth requires directing energy from the transponders at a sculpted parabolic reflector by an array of feedhorns (Fig. 4) controlled by a beam-forming network. The reflector then collimates the energy on Earth. This is a complicated, costly process, made even more so as customer requirements have grown. As an alternative, designers can use mathematical techniques to contour the otherwise smooth parabolic reflector, so that it can produce the desired pattern when fed by only a single horn. The reflector surface is mechanically shaped and creates a rippling or dimpled appearance. The dimples are precisely positioned by extensive computation to produce the desired pattern, thereby supplanting

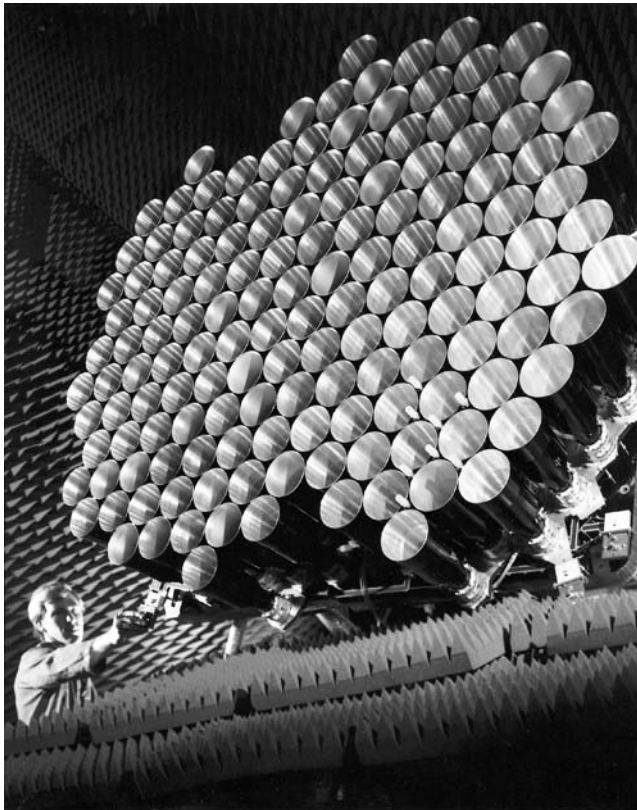


Figure 4. Feedhorn array that helps generate spot beams for INTELSAT VI satellite is checked out in simulation laboratory (courtesy Hughes Electronics Corp.). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

complex multiple feed arrays for directing radiated energy in the desired pattern. Elimination of the feed arrays reduces weight and expense.

There is increasing use of reconfigurable spot beams, narrowly shaped beams that illuminate specific areas, thereby increasing the power concentrated in those areas. By using onboard digital processing, modern satellites can generate many, even hundreds, of reconfigurable beams simultaneously. This has created the opportunity for satellite operators to obtain greater bandwidth by reusing the limited amounts of assigned spectrum. Using the same frequencies a number of times does not create interference if these frequencies are confined within spot beams directed at geographically separated regions (34).

Special Challenges: Attenuation and Echo Cancellation. Attenuation of electromagnetic energy becomes increasingly severe at higher frequencies above 12 GHz, as signals are absorbed by the atmosphere. Some spectral points experience more severe effects than others (35). At 22 GHz, energy is almost totally absorbed by water vapor, and at 60 GHz by oxygen, rendering these and some other higher frequencies useless for communications between satellites and Earth. Attenuation of microwave signals by rain droplets in the atmosphere can become a severe problem at higher frequencies because of the relatively large size of the droplets compared to the diminishing size of wavelengths at these frequencies (36). The droplets absorb and scatter the microwave energy. Rain attenuation in the Ku band can account for about 2 dB in signal loss and three to four times that amount in the Ka band. The presence of rain attenuation is not accounted for in the power balance equation because of the random nature and unpredictable magnitude of rain cells. Systems have to be designed with sufficient link margin to accommodate rain attenuation. This could mean increasing transmitter power and/or the gain of the transmitting antenna. The amount of attenuation will vary with geographic area and seasonal rain activity. A system designed for 99.5% availability—where attenuation is less than 2 dB in 99.5% of the time—could experience about 4 hours of outage in the months of heaviest rainfall. Under these conditions, a viewer's TV picture could freeze, and the sound would stop. A large telecommunications installation typically would need a 6- to 8-dB link margin to be certain that it could offset the effects of random rainstorms. Or the attenuation could be averted by using spatial diversity, transmitting from the satellite to two widely separated Earth stations (37).

Despite rain attenuation in space-to-Earth or Earth-to-space links, Ka-band frequencies are successfully used in space-to-space communications as intersatellite links. As cross-links for communicating between satellites, they are free from rain attenuation. Such links are used by some military satellites and new commercial personal mobile communications satellite systems. Intersatellite cross-links are transponder links that transmit and receive information between satellites, thereby establishing connectivity among satellites in a system. In military communications applications, cross-links permit authorities to pass information securely from one GEO satellite to another in achieving global or near-global connectivity without fear that potential enemies will eavesdrop on or jam these links. In the commercial satellite-based personal mobile communications systems coming on-line, cross-links are a key to providing global coverage. For example, each satellite in the system's six orbital rings is equipped with four Ka-band (23.18–23.38 GHz) antennae, two fixed and two gimbaled. A satellite can

communicate directly with the satellite before and after it in the same orbit through the fixed antennae. The gimballed antennae on a satellite allow communicating with the two neighboring satellites on either side in adjacent orbital planes. A user's call can be relayed via the cross-link from satellite to satellite until it reaches a satellite within reach of its destination. At that point, calls are downlinked directly to the desired parties, if they are system subscribers, or to local gateways and through the public switched telephone network to the intended individuals. Ka-band frequencies are expected to have widespread applications in broadband, high-power commercial satellite systems in the future.

Digital echo cancellers have been developed and installed in digital phone exchanges and satellite networks to eliminate the once-troubling problem of voice echoes on satellite-relayed telephone calls (38). The echoes were produced by the reflection of a party's speech created by a hybrid in the telephone network. The hybrid's function is to route energy between the two-wire phone pair connecting a caller in an office or at home to the four-wire trunk line on long-distance calls. Additional hybrids that may be present between the two parties in the network add to the problem. The difficulty is further compounded in a GEO satellite connection by the nearly 36,000-km separation between the speakers and the satellite. The round-trip propagative time for electromagnetic energy across the lengthy distance creates a 260-ms signal delay. In a satellite hookup, telephone parties could be disturbed by echoes of their own voices delayed by 260 ms in the midst of phone conversations. Digital echo cancellers in the form of software or firmware solve the echo problem by sampling a digitized version of speech and mathematically removing the echo. The echo is suppressed by a factor of 10,000 (39). Though the echo is essentially gone, the delay caused by the signal transit time to and from a GEO satellite is not. Some people are annoyed by the delay, but various studies indicate that 90% of telephone subscribers have no serious objection to it.

Ground Antennas (40). As satellites have grown in size and power, the size of Earth stations for accessing a satellite's communications capacity has dramatically declined, giving users greater flexibility, easier access, and lower costs. The antennae of large communications facilities today are about 3–9 meters (9.8–29 ft) in diameter for operation in the C and/or Ku band. At Earth stations, terrestrial signals typically are multiplexed, encoded, and modulated in the baseband section of the ground terminal. The signals are upconverted and amplified before transmission through the antenna to a satellite. Weak signals downlinked from the satellite are amplified through a low-noise amplifier, downconverted, and passed to baseband equipment for demodulation, processing, and interfacing with the terrestrial networks.

Customer premises antennae for offices or homes measure about a meter (3 ft) or less in diameter. Individual subscribers can receive TV programming broadcast to their homes by satellites operating in the Ku band through antennae as small as one-half meter (18 in) in diameter for some systems (e.g., Hughes' DIRECTV—see discussion in the article on Commercial Applications of Communications Satellite Technology). Subscribers to satellite-based personal mobile communications systems (e.g., Globalstar, Thuraya, and ICO) can speak to other parties via L-band satellites through handheld, cellular-like phones that have small omnidirectional antennae. The appearance of small, compact ground

terminals, called very small aperture terminals (VSATs), has led to the widespread use of satellite networks by businesses, schools, governments, and telecommunications organizations throughout the world. Typical VSATs vary in size, depending on application, but generally range from 0.85 m (2.8 ft) to 1.2 m (3.9 ft) in width or diameter for Ku-band systems.

Arranged in what is called “star,” or “hub-and-spoke,” network architecture, VSATs provide multipoint communications between a hub station at an entity’s headquarters or data center and a multiple number of remote sites. A computer at the hub handles switching for voice or video. The hub station broadcasts data via satellite to the VSATs at various remote sites. The terminals of individual remote sites are smaller and less powerful than the higher power, more expensive one at the hub. The remote sites can complete a return link to the hub through the satellite, thereby bypassing any terrestrial link, but their signals are too weak to be received by, or to interfere with, other sites in the network. (More expensive, “mesh” networks enable all sites to communicate directly with each other.) The “star” concept enables users to balance the higher cost of the hub against that of multiple lower cost VSATs. Manufacturers have reduced the size and weight of VSATs and increased their capability by adding more processing power. Size reductions help in two respects: the smaller the size, the lower the cost of installation on a customer’s premises, and the smaller the area occupied by the VSAT.

For all satellite applications, satellite systems are controlled and monitored from the ground by skilled technical personnel through one or more tracking, telemetry, and command (TT&C) stations. INTELSAT, for example, operates six such sites internationally to ensure reliable, line-of-sight contact with its entire orbital fleet (41). The exact positions of the satellite are determined from data secured by on-site tracking antennae. Personnel, consoles, and data processing equipment are housed in a satellite control center, which may or may not be collocated with a TT&C station. The satellite control center is manned and operated 24 hours a day throughout the year.

Spacecraft Design. Communications satellites have two fundamentally different designs. One is the spin-stabilized satellite, or spinner; the other is the three-axis, or body-stabilized satellite (42). The spin-stabilized satellite has a cylindrically shaped body that is spun about its axis, typically at a rate of 60 rpm to achieve stability inertially, much like a toy top becomes gyroscopically stable by spinning. The attitude and orientation of the satellite can be altered by a number of onboard thrusters. Almost the entire satellite spins, with the notable exception of its antennae, which are despun and pointed in a fixed direction toward Earth. Subsequent innovations in this concept led to a dual-spin design in which the entire payload, not just the antennae, is despun. To achieve stability, the body-stabilized design relies either on an internal gyro, called a “momentum wheel,” or a set of reaction wheels. The satellite’s control system compensates for changes in attitude by applying small forces to the spacecraft’s body.

The spinner design is the simpler of the two, and that simplicity translates into relatively low cost and long life without much ground intervention. The spinner employs a novel control system to maintain satellite attitude, or orientation, in space. A single axial thruster pulsed in synchrony with signals from an

onboard Sun sensor controls the attitude of the satellite's spin axis. When that thruster is operated in a continuous mode, it provides velocity control around the spin axis. Pulsed radial jets control velocity in a particular direction. The control concept is covered by what has become a seminal patent, referred to by the name of its inventor, Donald Williams, and validated in extensive legal proceedings. Prime power for the satellite comes from solar cells that cover the circumference of the spinning cylinder. More and more power has been squeezed from this design by enlarging the cylinder and by such steps as adding a second concentric solar-cell-covered cylinder that telescopes out from the first in space. But covering the spinning cylindrical surface with solar cells means that, at any given time, the Sun is illuminating no more than a third of the surface solar cells. The remaining two-thirds are not converting solar energy into electrical energy. This relatively inefficient use of energy resources becomes less acceptable when higher satellite power levels are necessary. As the power required exceeds 1 or 2 kW, the body-stabilized configuration becomes the preferred choice.

Because the three-axis stabilized spacecraft does not rely on a spinning body for stabilization, the satellite can assume any convenient shape, usually a box, for the convenience of its communications function. One surface of the box, on which antennae are mounted, is oriented toward Earth. Solar cells are mounted on flat panels that extend in wings from either end of the body and are oriented toward the Sun. The solar panels can be kept pointing at the Sun by motors aboard the satellite, as it revolves about Earth.

In addition to more efficient use of its solar cell resources and its potential for higher power levels, the body-stabilized satellite provides better pointing accuracy. This permits more accurate pointing of the satellite's antenna beams. Otherwise, the size of the antennae would have to be enlarged to ensure sufficient gain across the coverage area. When its solar wings and antennae are folded in a stowed position, the body-stabilized configuration lends itself to more efficient use of the volume within the shroud of a launch vehicle. And the box-like shape has more surface space for mounting antennae, permitting designers to employ more complex antennae. The drawbacks of the three-axis design stem from its relative complexity. Its momentum wheels have to be controlled in speed and pivoting. The satellite has to be commanded frequently, requiring an on-board computer. A typical spin-stabilized communication satellite can now be constructed in about 60% of the time required to produce a body-stabilized version. The difference in time is a measure of both the greater capability of the latter and the greater simplicity of the former.

Spacecraft Subsystems. An active satellite consists essentially of two parts—a payload and a bus. (A passive satellite is one that does not generate energy; an example is the Mylar-coated reflective balloons considered promising candidates for communications relay in the early 1960s. The payload contains the satellite's communications equipment and antennae that create an infrastructure for communicating with users throughout a continent or in regions or countries where service is supplied. The bus has the task of protecting the payload during the demanding launch period, placing the payload into its assigned orbit or orbital slot, and maintaining it there. The bus supports and maintains the payload throughout its lifetime (43).

A communications satellite contains seven subsystems (44):

- The communications subsystem (“payload”) contains the satellite’s radio-frequency equipment. A wideband receiver at the front end of the subsystem accepts incoming communications channels that occupy a specified band of frequencies. Then the channels are separated according to frequency by a multiplexer, or bank of filters, and apportioned among the payload’s various transponders. After amplification in the transponders, the channels are recombined by another multiplexer for retransmission to the ground.
- The power subsystem generates, regulates, and controls power obtained from the solar arrays and onboard batteries primarily for use by the communications payload. This subsystem also maintains operation of the satellite during periodic solar eclipses.
- The attitude control subsystem senses any deviations from proper pointing directions and keeps the spacecraft and the antennae pointing in the correct directions—the solar arrays pointing toward the Sun and the radiators away from the Sun.
- The propulsion subsystem generates thrust to place a GEO satellite into a desired orbital slot and to adjust its position periodically to offset movements in the (1) north–south direction due to solar and lunar gravitational attraction and (2) east–west direction due to the oblateness of Earth’s poles. The last-named function is called stationkeeping. GEO satellites contain either a solid rocket apogee kick motor (AKM) or a liquid bipropellant (separate fuel and oxidizer) system. The function of the AKM (45) and in part that of the bipropellant system is to insert the satellite into geosynchronous orbit when it reaches the apogee of a geosynchronous transfer orbit. The satellite is placed in an elliptical transfer orbit by a perigee kick motor during the final phase of the launch sequence. Besides performing the apogee kick function, the bipropellant system also helps raise the perigee of the transfer orbit to coincide with its apogee in geosynchronous orbit, a process called orbit raising. It also handles the stationkeeping duties. On satellites that have an AKM, a monopropellant system performs orbital positioning and stationkeeping duties. The tankage, valves, lines, thrusters, and fuel of this subsystem account for a significant portion of the mass of a satellite at launch and even after initial insertion into GEO orbit.
- The thermal control subsystem radiates the heat generated onboard the satellite into space. The thermal environment inside the satellite is kept at room temperature and all excess heat has to be radiated from the satellite. For this purpose, the subsystem uses such devices as thermal blankets and reflective mirrors. If the satellite has an AKM, as spinners do, an insulating wall and thermal barriers protect components from heat generated by the motor firing. New satellite designs are adding more TWTAs that have higher power outputs to their payloads. Despite increases in TWTA efficiency to as much as 70%, the remaining 30% is generated as useless heat that must be removed.
- The TT&C subsystem enables ground personnel to monitor the health and status of the satellite and issue commands to the satellite. It telemeters to

the terrestrial TT&C station information regarding satellite temperatures, remaining fuel, TWTAs performance, and pointing directions. The command portion accepts signals from the ground for controlling housekeeping functions, recharging batteries during solar eclipses, and dumping the energy buildup from the momentum wheels of a body-stabilized model. When the satellite comes to the end of its mission life as onboard fuel nears depletion, it can be commanded through the TT&C subsystem to deorbit and turn off its communications subsystem.

- The structures subsystem is the chassis that provides physical support and protection for sensitive equipment during launch when the satellite must survive the effects of severe acoustic and vibrational forces. It employs a truncated cone or trusswork with panels for mounting bus and payload electronics.

Reliability, Lifetime, and Cost. Satellite designers constantly seek to improve reliability so that service delivery by communications satellites is ensured throughout their contracted mission lives. Satellites have to operate in space without physical intervention for as long as 15 years in the case of newer GEO models. And the space environment is unforgiving; such phenomena as ionizing radiation pose hazards for sensitive electronics. Satellites are usually designed with redundancy for all critical components to prevent catastrophic failures. The increased reliability achieved through the years is attributable to numerous factors. Among them are better designs, fewer parts, improved manufacturing processes, and fewer electronic interconnections due to the application of semiconductor chips that have high circuit density. Potential life-limiting spacecraft elements are batteries, lubricants, thrusters, and TWTAs. Mechanical deployments of antennae and solar panels are another source of possible difficulties. Once a matter of great concern, TWTA technology and manufacturing know-how have advanced to the point where the chances of losing a transponder because of a TWTA failure are slim. Like other portions of a satellite, transponders are designed with redundancy to permit continued operation in the event of a failure. Now, transponders need only a single TWTA to back up three or four active TWTAs, not the one for two ratios of two decades ago. The TWTA suffers from only a single wear-out mechanism—its cathode. But the design of this element is well proven by millions of hours of actual operating experience. The infrequent failures that occur result from processing errors, not design inadequacies. This compels the tube manufacturer to maintain constant vigilance over production processes.

Commercial GEO satellites are actually designed for a lifetime as long as 18–20 years, but this life span, or design life, is limited by practical matters, such as the inability to replace degraded components and battery cells in an orbiting satellite. The satellite's design life is also limited by the onboard fuel supply. Therefore, the actual mission, or operational, life, of necessity, is shorter than the design life—no more than 12 to 15 years. Satellites in LEO orbits have shorter operational lifetimes—perhaps 5 to 8 years—limited as they are at very low altitudes by the effects of atmospheric drag and radiation. The contracts of GEO satellite manufacturers with their customers generally contain performance

incentives that reward the builder for achieving specific operational lifetimes. Just as Early Bird surprised its owners by unanticipated longevity, many other commercial satellites have as well. For example, Marisat 2, one of INMARSAT's satellites for mobile communications, had a 5-year mission life when placed in orbit in 1976. Twenty-two years later it was still providing service for the international maritime consortium, albeit from an inclined geosynchronous orbit.

Buying a satellite system involves a large investment by a communications entity. For instance, one major mobile communications satellite system, Thuraya, that was built in the late 1990s has an estimated value of \$1 billion. The figure includes the cost of two GEO satellites, the launch of one, ground facilities, training, and user equipment. (See further discussion of Thuraya in the article on Commercial Applications of Communications Satellite Technology.) The customer's cost of communications satellites, as measured by several key parameters, has been declining as a consequence of enhancements in satellite power, bandwidth, and lifetime. Thus, the satellite's price per kilowatt of power, price per transponder, and price per transponder year are all declining.

The costs of launching a satellite, the associated launch services, and insuring the launch and in-orbit satellite performance are a significant portion of the total price of acquiring and orbiting a satellite. The cost of insurance varies with the satellite, launch vehicle, recent claim experience, and available underwriting capacity, so that it varies to reflect current circumstances. During a period of 20 years, launch insurance premiums have varied from 6–20% of the insured value; the average is in the 10–13% range. In general, acquisition of the satellite represents slightly less than half the cost of a satellite in orbit; launch services and launch insurance account for the remainder.

New Technologies

Digitalization. The latest communications satellites incorporate greater amounts of digitalization to enhance their capabilities and flexibility (46). Digital signal processors of increasing capacity began to appear in satellite payloads during the 1990s. They can perform such roles as routing signals to any one of scores of individual spot beams and controlling and forming the beams. The beams are generated by a planar phased array antenna that contains multiple elements. In the process, called digital beam-forming, each element of the array captures a portion of the signal. The processor computes and controls the relative phase and amplitude of each element and can electrically bend the antenna beams into any desired shape. In this digital beam-forming antenna, the outputs from the array are sampled by an analog-to-digital converter and are stored in a processor that computes the phase changes needed to bend the antenna beam into the desired shape. A nearly unlimited number of these virtual antennae can be created. Each of them, shaped differently and pointed in a different direction, emanates from the one multielement phased array. They can be repointed quickly because the process of generating them is mathematical and there is no movement of mass involved in repointing. The ability of the processor to generate as many as 200 or 300 spot beams, to reuse limited frequencies extensively, and to support thousands of voice channels simultaneously makes possible such new

satellite services as personal mobile telephony, broadband data exchange, and multimedia communications. Digital processors give operators the freedom to satisfy changing customer demands by reconfiguring the power level, frequency, and beam shape after a satellite is in orbit (47).

Four separate generations of digital processors evolved in rapid sequence during the initial 10 years after they were introduced into communication satellites. They demonstrate successive improvements in circuit density, decreasing power consumption, and compactness that led to a progression in capability. The first generation, for example, for a military payload, used 13 different types of application-specific integrated circuits (ASICs); each circuit had about 15,000 gates. By the third generation, for application in a personal mobile communication satellite (Thuraya), the number of ASIC designs dropped to nine, but their density rose to 1.5 to 3.5 million gates per ASIC—a 100-factor jump over the first-generation level. By the fourth generation, for application in a broadband, multimedia satellite (Hughes's Spaceway, see discussion under "Commercial Applications of Communications Satellite Technology"), the ASICs count climbed to the four-to-seven million-gate range. Processor power consumption declined by approximately 50% from generation to generation. Consequently, the power required by the fourth generation was merely 13% of that of the first. The fourth-generation design provides 960 times the processing capability of the first, measured in raw processing power, or number of tera operations per second.

In a satellite application, a digital processor functions as a massively parallel digital computer with analog inputs and outputs. It converts analog radio signals received from individual terminals into numbers, performs the necessary computations entirely in the mathematical domain, and reconverts the resulting data into analog signals for retransmission to the terminals. This is called a "demod/remod" design. A demod/remod digital processor also performs filtering, switching, demodulation, modulation, beam-forming, and other functions without introducing distortion or drift. It can correct errors, reconstruct corrupted signals, and eliminate noise. The demod/remod design relies on known characteristics of the user's signals. An alternate design, the transponded system, knows nothing about the format of incoming signals and is insensitive to changes in them. Processors of this design perform little or no filtering. Also, they cannot flexibly allocate downlink power to accommodate different sizes of user terminals.

Power. Several technical innovations have contributed to the ongoing trend toward increased satellite power. The introduction of gallium arsenide solar cells marked a turning point in efforts to improve the efficiency with which photovoltaic cells convert solar light into electrical energy. Gallium arsenide cells achieved efficiencies of 21.6%, nearly doubling the 12.3% figure obtained from traditional silicon cells. The increase enabled designers almost to double the dc power output from solar arrays of comparable size and mass. The greater power can be translated into an increase in the capacity of the communications payload. Alternatively, when the extra power is not needed, the size and mass of the satellite solar cell arrays can be reduced, thereby decreasing satellite mass and possibly launch vehicle cost.

The initial gallium arsenide cells were dual-junction devices that have two semiconductor junctions—a gallium arsenide layer on a single crystal

germanium substrate and a gallium indium phosphide layer atop the gallium arsenide. The improved efficiency stems from the ability of each layer to convert a different part of the light spectrum into electrical power. When the Sun's rays strike the top layer of the cell, shorter wavelengths are converted to power. The top layer is transparent to the longer wavelengths, which penetrate the gallium indium phosphide layer and strike the gallium arsenide layer, where they too are converted to electrical power. The cells are grown in an epitaxial chamber by a gas reduction process. Continuing work has led to triple-junction gallium arsenide cells. These enhanced the efficiency of earlier gallium arsenide cells by 20%, and brought conversion efficiency up to 26.8%. Four-junction cells are expected to reward satellite manufacturers with still higher solar cell efficiencies in the early years of the new millennium. The solar arrays of the new body-stabilized Hughes HS 702 satellite employ gallium arsenide cells, as did several other satellites in the late 1990s. Gallium arsenide cells are an important contributor to the satellite's ability to generate as much as 15 kW of end-of-life (EOL) power. Satellite builders generally specify power levels expected at the end of the satellite's lifetime, as opposed to higher power levels at the beginning of life (BOL). The difference reflects anticipated degradation in the solar cell output due to the deleterious effects of X rays, solar protons and electrons, and other particles encountered in different orbits.

Innovative solar concentrators mounted along the satellite's solar wings add significantly to the power generated by the HS 702 satellite's solar cell panels. The angled solar reflector panels add 50% to the power output from the solar cell panels on the wings by concentrating more solar energy on the solar cells. The reflector panels are angled outward along both sides of the wings to form a shallow trough, and the solar panels are at the bottom. Sunlight that otherwise would not strike the flat panels is reflected back from the reflectors onto the solar panels. The HS 702 is configured to enable designers to tailor the power output from the arrays to satisfy specific customer requirements. This is done by choosing any one of six different solar array configurations that have up to five panels per wing.

Power subsystems of today's communications satellites rely primarily on nickel-hydrogen batteries for powering payloads during eclipses. But batteries are heavy and costly; they weigh of the order of 500 kg. The next battery technology expected to provide capacity for higher power satellites early in the new century was in the development stage in the closing days of the 1990s. Lithium ion batteries hold the promise of halving the mass of nickel-hydrogen satellite cells while offering a less expensive alternative. One possible version built of plastics contains a nonliquid electrolyte within the plastic. High-speed flywheels that extract the energy of a spinning mass are another promising technology but are further removed from application (48).

Propulsion. Beginning in the early 1960s, cesium and mercury electric ion propulsion systems flew aboard U.S. Air Force and NASA satellites. Since then, the performance of new commercial communications satellites has been getting a dramatic lift from a newer spacecraft propulsion system first flown on a non-commercial European Space Agency satellite in 1992. Also a form of electric propulsion (49), the xenon ion propulsion system (XIPS) makes possible reductions of as much as 90% in the mass of a satellite's fuel (Fig. 5).



Figure 5. Thruster for xenon ion propulsion system (XIPS) protrudes from PanAmSat PAS-5 satellite. The electric propulsion system complements the satellite's less efficient bipropellant propulsion system to reduce substantially the amount of fuel required by that system (courtesy Hughes Electronics Corp.). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Normally, at the beginning of orbital life, chemical propellants required by a satellite's bipropellant propulsion system for attitude control and stationkeeping throughout the mission account for at least 25% of satellite mass. This could amount to 500 to 700 kg (1102 to 1544 lb) for a conventional communications satellite. The comparable amount of fuel for the more efficient XIPS propulsion system would be only 75 to 105 kg (165 to 232 lb) of inert xenon gas. At a cost of about \$30,000 per kilogram to place that mass in orbit, millions of dollars could be cut from launch costs. The mass reduction can also be translated into an increase in satellite payload, or communications capacity. Or, with additional xenon, a longer satellite mission life can be obtained because the quantity of onboard fuel is the principal limiting factor on the satellite's operational life. Alternatively, the XIPS technology could produce benefits that combine these savings. A XIPS-equipped satellite uses the impulse generated by a thruster assembly and ejects electrically charged particles at high velocities. The entire system consists of a source of pressurized xenon propellant, a power processor, and the cylindrically shaped thruster assembly. Thrust is created by accelerating

positive ions generated within the assembly's ionization chamber through a series of gridded electrodes at one end of the assembly. The electrodes create more than 3000 tiny beams of thrust. The beams are prevented from being electrically attracted back to the thruster by an external electron-emitting device called a neutralizer. XIPS generates a very high specific impulse, expressed in seconds as the ratio of thrust to the rate at which the propellant is consumed. The higher the specific impulse, the less propellant required. The specific impulse of a XIPS system is 2600 to 3800 seconds, depending on the satellite configuration—roughly 10 times that of a bipropellant system. (Note that the 1964 NASA SERT I satellite, which used a combination of mercury and cesium ion, had a specific impulse of about 5000 seconds, but this was not proven sustainable or available all the time.)

The thrust level of a modern XIPS system is very small. For an HS 702 satellite, it is 0.165 newton, or 0.037 pound of force. In a satellite such as the HS 702, the XIPS system can also augment the satellite's bipropellant propulsion system in orbit-raising. Using XIPS to help circularize the elliptical transfer orbit further reduces the amount of chemical propellant needed on the satellite, thereby again shrinking satellite mass. The extra mass reduction gives customers even more latitude in choosing among the benefits of a less expensive launch vehicle, additional payload, or longer mission life. The HS 702 XIPS is designed for a maximum orbit-raising duration of 90 days, compared to 3 days for the higher thrust but less efficient bipropellant system. The trade-off for satellite customers is one of mass reduction benefits versus delayed arrival on station.

Systems. The power per kilogram of satellite mass has grown steadily. Early in the 1980s, when the current generation of spinners was introduced, the figure stood at about 1 W per kilogram. It climbed to 2 W per kilogram for the HS 702's predecessor, the three-axis HS 601, in the early 1990s. Later in the decade, the figure reached 4 W per kilogram for the HS 702. It is likely to rise for later generation satellites. The availability of higher power from a satellite like the HS 702 enables operators to increase transponder effective isotropic radiated power (EIRP), a measure of performance that takes into account transmitter power, antenna gain, and the losses in the transmitter waveguide. This is the satellite portion of the previously cited power balance equation. A high EIRP permits customers to use smaller Earth antennae such as the 45.7-cm (18-in) and 60-cm (24-in) home antennae for direct-to-home TV in the United States and Europe, respectively. Or operators can apportion available power among a larger number of transponders, effectively doubling the transponder complement and reducing cost per transponder in orbit. Still another option is to choose a combination of both performance benefits.

The newer, higher power satellites are the backbone of high-speed, broadband satellite services (e.g., Teledesic, Spaceway, Astrolink—see later) coming on line early in the new millennium. Because switching functions are handled by digital processors in the satellite rather than on the ground, whole new vistas will be opened for corporate and other VSAT users. The small terminals will be able to broadcast at data rates of multiple megabits per second, compared to the recent 128-kbps rate. And hundreds of megabits per second will be downlinked to them.

Digital compression techniques account for a panoply of new satellite-delivered entertainment and business offerings. Digital video compression

effectively increases the use of communications satellite capacity by factors of 4–8, and could double that by the early 2000s. Thus, a 16-transponder satellite could transmit 256 TV channels. The doubling could reduce an operator's operating and capital costs by an order of magnitude. Before the recent application of compression techniques, only one analog broadcast-quality video signal could be transmitted through each transponder on a satellite. Consequently, the number of video programming channels was limited by the transponder count. Now, using digital compression, broadcasters can convert analog video channels into digital data. The digital signals are compressed and transmitted by the satellite to end users. At the end user's set-top receiver, they are decompressed and converted back into the customary analog form for viewing. Each video channel occupies only a portion of the transponder's bandwidth, so additional channels can be squeezed into that transponder. The compression process involves representing video signals by a series of numbers, eliminating redundancy in the scenes, retaining only frame-to-frame differences, and reconstructing the video at the receiving end from the sequence of numbers.

A comparable transformation is occurring in the transmission of CD-quality audio by satellite. The entire contents of a compact disk amounting to 620 megabits can be compressed into 32 megabits and replaced by a "flash" random access memory (RAM) chip. The contents of the flash chip can be downloaded in 1.5 s from a satellite to a consumer by a direct-to-the-home transponder. There, it can be stored in the hard drive of a TV set-top box. A 10-gigabit hard drive has sufficient capacity to store the equivalent of approximately 300 CDs. Now the set-top box with speakers becomes the storage medium and solid-state "music player" for the home, replacing the stereo and compact disk player. Similarly, digitally compressed audio can now be transmitted by a satellite to a moving automobile through a "whip" antenna and stored in memory. Drivers can then select the music or audio programming they wish to hear. In another application of the same technology, pictures taken by a digital camera now can be sent around the world in seconds. A typical digital camera can store about 100 pictures on a flash chip. The chip can be removed and inserted into a computer to uplink the contents to a satellite for retransmission.

In recent years, there has been a resurgence of interest in both LEO and MEO orbits for communications satellites (50). The non-GEO orbits have won adherents among operators of global mobile personal communications systems. Those such as Globalstar, for example, permit subscribers who have handheld cellular phones to speak to others beyond the range of normal cellular transmission towers almost anywhere in the world. Note that the now-defunct Iridium handheld global mobile telephony system did not fail because of its LEO orbit, nor for any technical reason. Missteps in marketing and financing are generally cited. LEO orbits also are being used for two-way data communications and messaging systems (e.g., Orbcomm). All of these systems require that the user maintain line-of-sight contact with the satellite. Unlike terrestrial cellular phones, they have insufficient link margins to permit subscribers to place or receive calls from inside a building or other structure.

The interest in non-GEO orbits was stimulated by explosive growth in worldwide cellular phone usage and an accompanying upsurge in demand for greater telephony infrastructure. Designers of LEO systems rejected GEO orbits

on two counts (51). They believed that a GEO satellite system could not provide enough link margin for a satisfactory link between satellites and on-the-go mobile subscribers using small handheld phones that have omnidirectional antennae. They also maintained that the high latency, or 260-ms signal delay, from GEO orbit (plus any processing and speech compression delays) would be unacceptable to cellular phone subscribers. Other designers, however, concluded that an attractive compromise was offered by a MEO global mobile communications satellite system: global coverage that uses fewer satellites at high elevation angles; fewer handovers than a LEO system; and, at 100 ms, a shorter round-trip propagative delay than the 260-ms delay from geosynchronous orbit.

GEO mobile communications satellite builders resolved the margin issue by resorting to higher power satellites using very large deployable parabolic antennas of about 12.5 m (40 ft) in diameter (52). The combination of high spacecraft power and high gain of the large antenna ensures acceptable margins for satellite-based personal mobile communications. GEO systems of this type (e.g., Thuraya (53), and ACeS (54)—operated by PT Asia Cellular Satellite) offer mobile service to properly equipped subscribers anywhere in a region within the footprint of a single GEO satellite. But they do not offer global coverage. To do so would require additional satellites and, possibly, more than a single hop.

As for the latency issue, GEO satellite designers are convinced that users can accept as much as a 400-ms delay in voice communications. Consequently, GEO mobile systems are configured to keep the burdensome and unalterable round-trip propagative delay, plus compression and processing delays, within the 400-ms limit. At low orbital altitudes, LEO systems pay only a small penalty for shorter propagative delays, but they still must deal with appreciable processing, handover, and compression times, especially for geographically separated parties on long-distance calls.

Conclusions

Amidst the post-World War II Cold War environment, President John F. Kennedy urged all nations “to participate in a communications satellite system in the interest of world peace and closer brotherhood among people of the world.” By 1967, three INTELSAT satellites were operating over the Atlantic and Pacific oceans, providing ubiquitous global telephone and television communications in hundreds of languages. By the early to mid-1970s, new domestic satellite systems were created in many countries and regions, such as Canada, Europe, the United States, and Indonesia. By April 2000, the date this chapter was written, sophisticated communications satellites became a fundamental element of the world’s telecommunications infrastructure and an indispensable tool of the global marketplace. These satellites have enabled many new space-based applications to be introduced to businesses and consumers around the world. These include direct-to-home television broadcasting, handheld global mobile telephony, and high-speed broadband data exchange.

These sophisticated communications satellites have evolved since the mid-1960s. In 1965, the Early Bird (INTELSAT I) satellite weighed 76 pounds. The

mass of today's satellites can exceed 12,000 pounds. Strides in satellite technology have also extended the lifetime and transponder capacity by several orders of magnitude. The comparative sophistication of today's communications satellites derives from the application of several advanced technologies. These include digitalization, which is used for numerous functions to enhance their capabilities and flexibility. For example, by using onboard digital signal processing, modern satellites can generate hundreds of reconfigurable spot beams simultaneously. Among other things, this enables satellite operators to generate maximum revenue and to obtain greater bandwidth by reusing the limited amounts of assigned spectrum. Digital echo cancellers have also been developed and installed in digital phone exchanges and satellite networks to eliminate disturbing voice echoes on satellite-relayed telephone calls. Digital compression techniques have already increased satellite use capacity by a factor of 8, and soon a 16-transponder satellite might transmit 256 TV channels. This will greatly reduce an operator's operating and capital costs. Before the advent of digital compression, only one broadcast-quality analog video signal could be transmitted through each satellite transponder. Digitalization is also revolutionizing the transmission of CD-quality audio by satellite, thus enabling the start-up (to date) of three new satellite digital audio services: Sirius, XM Radio, and WorldSpace (see discussion under "Commercial Applications of Communication Satellite Technology").

Other advanced technologies are making their imprint on a satellite's power and propulsion. As gallium arsenide solar cells replaced traditional silicon cells, satellite power doubled and enabled satellites to generate as much as 15 kW of power at the end of life. As a replacement for mercury and cesium electric propulsion systems, xenon ion propulsion makes possible reductions of as much as 90% in the mass of a satellite's fuel requirements. At a cost of about \$30,000 per kilogram to place a satellite in orbit, millions of dollars can be saved in launch costs. This reduction can also be applied to an increase in satellite payload, longer satellite lifetime, or any combination.

These and all ongoing technological innovations are intended to provide maximum available bandwidth through frequency reuse; maximum flexibility for operators and users; lower launch and in-orbit operating costs; more space-based communications options for consumers; extended satellite lifetimes and greater reliability; larger satellite footprints, including those created by reconfigurable spot beams; higher power levels resulting from increased use of solar power; and reduced cost per transponder.

ACKNOWLEDGMENTS

Preparation of this article was made possible in large part through input from the following experts from Hughes Electronics or outside consulting organizations: Grant Beaston, Bernie Bienstock, Gene Cacciamani, Brian Clebowicz, Bruce Elbert, John Ellison, Barry Fagan, Mike Fitch, Laura Hajost, Bill Heathcote, Eugene Kopp, Joe Lore, Ruth Macy, Robert Mercer, Barry Miller, John Navak, John Perkins, Steve Pilcher, Lawrence Seidman, Stan Sosa, Kathy Sullivan, Brooker Thro, Emery Wilson, and Dieter Zemmrich.

BIBLIOGRAPHY

1. Gavaghan, H. *Something New Under the Sun: Satellites and the Beginning of the Space Age*. Springer-Verlag, New York, 1998, pp. 171–172.
2. Berry, F.C., Jr. *Inventing the Future*. Maxwell Macmillan, McLean, VA, 1993, pp. 35–36.
3. Hudson, H.E. *Communication Satellites: Their Development and Impact*. The Free Press, New York, 1990, Chap. 1.
4. Curtis, A.R. (ed.). *Space Satellite Handbook*, 3rd ed., Gulf Publishing, Houston, TX, 1994, pp. 1–4.
5. Reference 3, Chap. 2.
6. Reference 3, Chap. 2.
7. *Communications Satellites: 1958–1995*, Aerospace Corporation, El Segundo, CA, 1996, pp. 5–13.
8. Reference 2, pp. 37–40.
9. Reference 1, pp. 199–210.
10. Reference 1, pp. 171–198.
11. Reference 1, pp. 211–220.
12. Tedeschi, A.M. *Live Via Satellite*. Acropolis, Washington, DC, 1989. Covers the early history of COMSAT.
13. Reference 3, Chap. 3.
14. Reference 7, pp. 49–82.
15. www.intelsat.org
16. Six Centuries in Space. Supplement to *Via Satellite*, September 1992.
17. www.inmarsat.org
18. Six Centuries in Space. *Via Satellite*.
19. Six Centuries in Space. *Via Satellite*.
20. Reference 3, Chap. 12.
21. Morgan, T. (ed.). *Jane's Space Directory: 1998–1999*. Jane's Information Group, Surrey, England, 1998, pp. 53–93, 377–384.
22. Martin, J. *Communications Satellite Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1978, Chap. 3.
23. Pritchard, W., and J. Sciuli. *Satellite Communications Systems Engineering*. Prentice-Hall, Englewood Cliffs, NJ, 1986, Chap. 2.
24. Maral, G., and M. Bousquet. *Satellite Communications Systems*, 3rd ed., Wiley, West Sussex, England, pp. 273–338.
25. Pocha, U.J. *An Introduction to Mission Design for Geostationary Satellites*. Reidel, Dordrecht, The Netherlands, 1987.
26. Logsdon, T. *The Navstar Global Positioning System*. Van Nostrand Reinhold, New York, 1992, pp. 17–33.
27. LeVine, S. Small satellite phones: An expensive gamble. *The New York Times*, March 30, 1998, Sect. D, p. 5, col. 1.
28. Jamalipour, A. *Low Earth Orbital Satellites for Personal Communication Networks*. Artech House, Norwood, MA, 1998.
29. Gerding, B. *Telecommunications* 35–36 (February 1996).
30. Elbert, B.R. *The Satellite Communication Applications Handbook*. Artech House, Norwood, Mass, 1997, pp. 39–49.
31. Reference 21, pp. 124–163.
32. Elbert, B.R. *Introduction to Satellite Communication*, 2nd ed. Artech House, Norwood, MA, 1999, pp. 22–40.
33. Reference 23, Chap. 9.

34. Morgan, W.L., and G.D. Gordon. *Communications Satellite Handbook*. Wiley, New York, 1989, pp. 242–258.
35. Reference 23, pp. 173–181.
36. Reference 32, pp. 139–144.
37. *Rain Fade Compensation Alternatives for Ka-band Communication Satellites*. NASA, TM-107534, Washington, DC, 1997.
38. Reference 32, pp. 94–96.
39. Reference 23, pp. 367–382.
40. Reference 24, pp. 339–416.
41. *INTELSAT Annual Report*, Washington, DC, 1997.
42. Reference 24, pp. 539–548.
43. Chetty, P.R.K. *Satellite Technology and Its Applications*, 2nd ed. McGraw-Hill, New York, 1991, pp. 97–343.
44. Reference 23, Chap. 5.
45. Gibson, J. (ed.). *The Communications Handbook*. CRC, Boca Raton, FL, 1997, Sec. V, Art. 67.
46. Reference 24, pp. 110–139.
47. Reference 45, Sec. V, Art. 72.
48. Nelson, R.A. Spacecraft battery technology. *Via Satellite* 104–118, February 1999.
49. Reference 24, pp. 556–560.
50. Pattan, B. *Satellite-Based Cellular Communication*. McGraw-Hill, New York, 1997.
51. Miller, B. *IEEE Spectrum* 26–35 (March 1998).
52. Thomson, M.W. *Proc. 5th Int. Mobile Satellite Conf.*, June 1997, pp. 393–398.
53. Thuraya Satellite Telecommunications Company, corporate brochure, Abu Dhabi, United Arab Emirates, 1997.
54. Adiwoyo, A.R., co-workers, *Proc. 5th Int. Mobile Satellite Conf.*, June 1997, pp. 145–152.

STEVEN D. DORFMAN
Hughes Electronics
Los Angeles, California

CONVERSION OF MISSILES INTO SPACE LAUNCH VEHICLES

This article begins with a brief history of the Yuzhnoye Design Office, one of the leading design bureaus in the Soviet Union and Ukraine. The Yuzhnoye Design Office designed numerous strategic missile systems, launch vehicles, and spacecraft (1). One interesting aspect of this history involves the conversion of missiles into space launch vehicles by a team of developers led by the Yuzhnoye Design Office, well in advance of the officially announced USSR conversion effort.

Brief History

The M.K. Yangel' Yuzhnoye State Design Office (initially known as Special Design Bureau 586, abbreviated OKB-586) was established on 10 April 1954 in

Dnepropetrovsk, a city on the banks of the Dniepr River in central Ukraine. Mikhail Kuz'mich Yangel' was named General Designer. Before this, he had been Director of the Scientific Research Institute 88 (NII-88) in Podlipki (now known as Korolev), a city near Moscow. In 1946, NII-88 became the USSR's main center for missile development.

A group of missile specialists from the General Designer's Department of All-Union State Plant 586 (the former Dnepropetrovsk Motor Vehicle Plant, now the State Enterprise Production Association Yuzhnyi Machine-Building Plant) formed the core of the newly established OKB-586. In 1951, Plant 586 started mass production of the R-1, R-2, and R-5 missiles developed by NII-88 and the NII-88 Special Design Bureau 1 (OKB-1), which was headed by General Designer Sergei Pavlovich Korolev. OKB-586 and Plant 586 joined forces to establish a missile design and production center where everything was under one roof. A nationwide network of developers and manufacturers for the components, systems, intermediate products, and hardware specified for use in OKB-586 missile production and development activities was also set up along with OKB-586 itself.

The Soviet Government's intention was to have this newly established system of production and development facilities headed by OKB-586 become (relative to the existing system) a stronger, more productive scientific production cooperative for developing future USSR strategic missiles. The intent was to promote missile development and also to increase substantially the military effectiveness of the missiles themselves. Despite the success of NII-88 in developing the R-1, R-2, and R-5 missiles, high-level military and government personnel understood that these missiles and any others that could be developed using the same principles could not meet future strategic requirements. Each of these missiles had substantial deficiencies that prevented them from being used in real combat conditions.

The new team would eventually eliminate these deficiencies by developing future missiles within a completely new conceptual framework. This conceptual framework was based on several principles developed by NII-88 personnel in 1952 under the leadership of M.K. Yangel'. These were the most important of the principles:

- development of a series of missiles having ranges consistent with strategic requirements;
- deployment of these missiles in hardened silo launchers constructed for concealment from the enemy;
- development of these missiles to use high-boiling-point propellants that could be stored long term and avoid using the low-boiling-point propellants previously used that had poor storage qualities;
- the use of autonomous onboard control systems protected against enemy electronic countermeasures and avoiding subsystems that might be vulnerable to noise.

The design concepts based on the high-boiling-point AK-27I and TM-185 rocket propellants developed by OKB-1 and Plant 586 General Designer's Department personnel for the R-11 and R-12 missiles in 1952 confirmed that these

principles were valid and sensible. The establishment of OKB-586 and the appointment of M.K. Yangel' as its General Designer were evidence that the Government supported this new conceptual framework for developing new missile systems. This conceptual framework guided OKB-586 in all of its activities and was continually developed and enhanced, as various new development projects were implemented. In the process, OKB-586's main mission became developing strategic missile systems capable of inflicting a highly effective second strike against any aggressor if the Soviet Union were the target of a nuclear attack.

Despite the organizational issues that arose during the establishment of OKB-586, a lack of essential equipment and experienced personnel, insufficient research on high-boiling-point propellants, delays in developing the rocket engine and autonomous onboard control system, and delays in constructing the silo launcher, the first medium-range missile system, the 8K63, was placed into service less than 5 years after OKB-586 was established. Two additional missile systems, the 8K65 medium-range missile and the 8K64 intercontinental ballistic missile, were placed into service at 2-year intervals. For some time, these missiles served as the primary weaponry in the USSR Strategic Missile Forces arsenal.

Within this brief period, under the leadership of M.K. Yangel', OKB-586 had become the USSR's leading design bureau for developing the strategic missile systems that were most important to the USSR's defense capability.

From 1954 to 1991, a total of 29 strategic missile systems were developed by OKB-586 (from 1966 on, known as the Yuzhnoye Design Office) and the team of engineers under the leadership of Academicians M.K. Yangel' and V.F. Utkin. Thirteen of these systems were accepted for military service by the Strategic Missile Forces and became the backbone of their forces (Fig. 1).

Some of these systems have no peers in missile technology. Examples include the 8K69, 15A14, 15A18, the 15A18 M, the 15Zh60 fixed solid propellant missile system 15Zh60, and the 15Zh61 rail-mobile missile system, all of which played an important role in enabling the Soviet Union to reach strategic parity with the United States and in negotiations of the Strategic Arms Limitations Treaties between the Soviet Union and the United States.

The four generations of missiles accepted into Missile Forces armaments are shown in Fig. 2, together with the dates when they were placed into service. Each of these missiles had a unique purpose and unique characteristics, and



Figure 1. General Designers: S.P. Korolev (1907–1966), M.K. Yangel' (1911–1971), V.F. Utkin (1923–2000), S.N. Konyukhov (1937–).

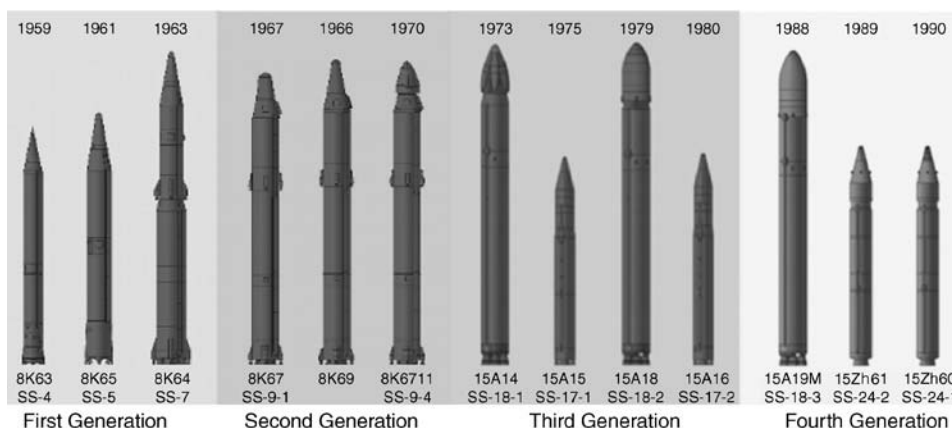


Figure 2. Military missiles developed by the Yuzhnoye State Design Office. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

there were substantial differences in specifications. They were all developed at different times, and each embodies the scientific, technical, and economic capabilities of the country at the time they were developed. In spite of these differences, however, it is possible to identify some trends typical of all four generations of missiles. The service life of the missiles tends to increase, the capabilities of the missile-defense countermeasures improve, the total energy output increases, the range increases and the operational specifications of the missiles all improve from one generation to the next.

The missiles of the fourth generation possess the highest military effectiveness. Whether in silos or during active flight, these missiles are able to preserve their performance in the face of any countermeasures. They are equipped with a very effective multifunctional antimissile system, which in combination with high survivability during active flight allows them to overcome with a high probability of success even a future adversary missile defense system. As of now the four types of the most highly developed military missiles 15A18, 15A18M, 15Zh60 and 15Zh61 are still deployed in the Russian Federation.

In addition to the missiles described, from 1957 on, OKB-586 also developed a variety of space launch vehicles. The most predominant idea involved developing a missile-based launch vehicle. This approach would lead to a substantial reduction in one-time and recurring costs, as well as a reduction in launch-vehicle development time due to the reduced amount of design and development work required and the ability to use the existing manufacturing infrastructure, the existing basic missile components available at the various manufacturing plants, and existing ground-based launch facilities.

This idea was implemented via the development of several launch vehicles based on the 8K63, 8K64, 8K66, 8K67, 8K68, 8K69, and 15A18 missiles; five of these launch vehicles—the 11K63 (Kosmos), 11K65 (Kosmos-2), 11K69 (Tsiklon-2), 11K68 (Tsiklon-3), and Dnepr—were in actual use. These five launch vehicles are the subject of this paper. Two additional launch vehicles, the Zenit-2 and Zenit-3SL, were developed without using military prototypes. By contrast with the missile-based launch vehicles, all stages of these launch vehicles used liquid

oxygen as the oxidizer and RG-1 kerosene as the fuel. A modified Zenit-2 first stage was used as a module in the Energia vehicle (Fig. 3).

The space ambitions of the Yuzhnoye Design Office were also embodied in the design and successful development of Module E, the lunar-lander portion of the lunar spacecraft developed as part of the lunar program. Since 1960, OKB-586 and the Yuzhnoye Design Office produced a series of research, commercial, applied scientific, and military spacecraft (more than 70 different types). This count includes the following Kosmos and Interkosmos spacecraft: AUOS, Okean, Taifun, Tselina, etc. Approximately 400 spacecraft designed by the Yuzhnoye Design Office and manufactured by the Yuzhnoye Machine-Building Plant have been launched into space, many of them aboard launch vehicles designed in-house (Fig. 4).

All design work at Yuzhnoye Design Office was performed under the direction of General Designer M.K. Yangel' before October 1971 and under General Designer V.F. Utkin between October 1971 and November 1990. M.K. Yangel' served as General Designer for the development of first-, second-, and third-generation missiles, but did not live to see the 15A14, 15A15, and 15Zh60 missiles certified for military use; however, these latter missiles were also based on his ideas, which he had to defend at many levels, up to and including the USSR Defense Council. M.K. Yangel' also attached a great deal of importance to space research. The 11K63 (Kosmos), 11K69 (Tsiklon-2), and 11K68 (Tsiklon-3) launch vehicles and Module E of the lunar spacecraft were developed under his leadership. At his initiative, a spacecraft design bureau was established at OKB-586 in 1960, spacecraft were produced at the Yuzhnoye Machine-Building Plant. Approximately three dozen types of spacecraft and several hundred spacecraft were launched under his leadership.

As Yuzhnoye Design Bureau Chief Designer, Vladimir Fedorovich Utkin had enormous influence on the development and delivery of the third- and

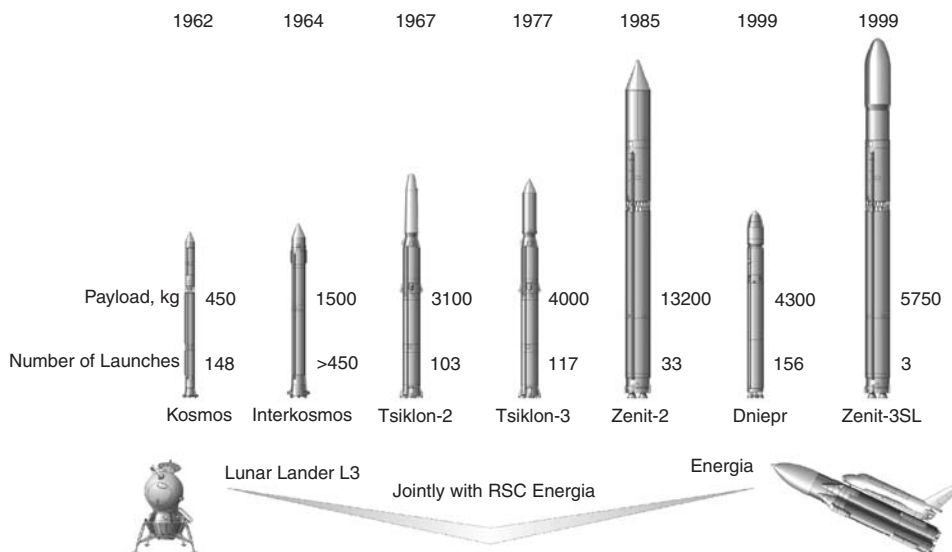


Figure 3. Launch vehicles developed by the Yuzhnoye State Design Office. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

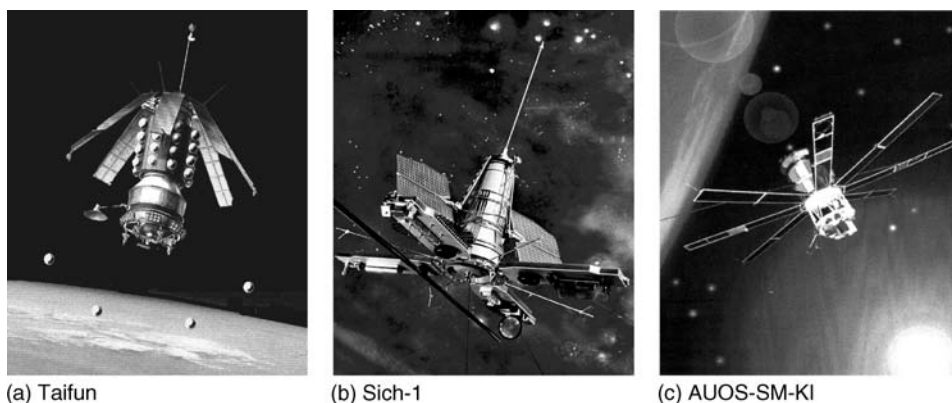


Figure 4. Spacecraft developed by Yuzhnoye State Design Office. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

fourth-generation missiles, as well as the Tsiklon-3 and Zenit-2 launch vehicles. The Energia launch-vehicle module unit and approximately 40 other types of spacecraft were developed under his leadership. The highly efficient and reliable Zenit-2 launch vehicle served as the basis for developing the Zenit-3SL Integrated Launch Vehicle core of the offshore launch platform developed under the Sea Launch program. Work on the Sea Launch project began on 25 November 1993 when an agreement was executed between the American aircraft and missile company Boeing, the Russian Rocket and Space Corporation Energia, the Norwegian company Kvarner, and two Ukrainian enterprises—the Yuzhnoye State Design Office and the State Enterprise Production Association Yuzhnyi Machine-Building Plant. Geosynchronous satellite launches via Sea Launch began on 28 March 1999 with a demonstration launch of the Zenit-3SL launch vehicle.

As before, development of future launch vehicles remains at the center of attention. A more powerful launch vehicle, the Tsiklon-4, has been developed on the basis of the Tsiklon-3. Work is currently underway on the Air Launch and Mayak projects, and various approaches for modernization of the Tsiklon-2, Zenit-2, and Dnepr launch vehicles are also being explored.

From Missiles to Launch Vehicles—the First Missile (8K63) and Launch Vehicle (11K63)

8K63 (SS-4) Missile (Fig. 5). The first strategic missile to embody the new concept developed under the leadership of M.K. Yangel' was the 8K63 (SS-4). The Government task order for developing the missile was issued to coincide with establishment of the OKB-586 design bureau.

The major responsibilities for development of the 8K63 missile/missile system were allocated as follows:

- OKB-586, Chief Designer M.K. Yangel'—systems engineering of missile and missile system as a whole;

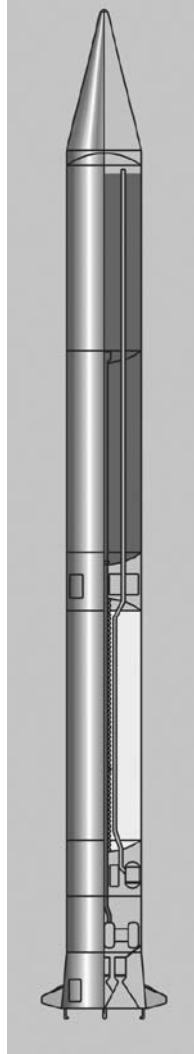


Figure 5. 8K63 missile. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

- KB-11, Chief Design Engineer S.G. Kocharyants—design of warhead and related equipment;
- NII-885, Chief Design Engineer N.A. Pilyugin—design of the autonomous onboard control system;
- NII-944, Chief Design Engineer V.I. Kuznetsov—design of gyroscopic instruments;
- OKB-456, Chief Design Engineer V.P. Glushko—design of RD-214 engines;
- Spetsmash State Special Design Bureau, Chief Design Engineer V.P. Barmin—design of aboveground and silo-based launch facilities.

These chief design engineers became the most active proponents of Yangel's approach to missile system development.

The 8K63 consisted of a monocoque single stage that had a nose cone section for the nuclear warhead, cylindrical fuel tanks, an instrument section that had an autonomous onboard control system, and a conical tail compartment containing a fixed RD-214 four-chamber engine using TM-85/AK-27I high-boiling-point propellants. The AK-27I oxidizer is an iodine-inhibited mixture of nitric acid (70%) and nitrogen tetroxide (27%), and TM-185 is a modified-kerosene hydrocarbon fuel. At the time, these were the best-known high-boiling-point propellants that had an adequate production infrastructure.

The missile had a launch weight of 41.7 metric tons, a length of 22.1 m, and a body diameter of 1.652 m. Its RD-214 engine produced a thrust of 648/744 kN and a specific impulse of 2300/2640 N · s/kg (sea level/vacuum). TG-02, a xylidine/triethylamine mixture, was used as an ignition propellant to ignite the fuel in the RD-214 combustion chamber. The engine turbopumps were operated using a gas-generator mixture obtained by decomposing hydrogen peroxide in the presence of potassium permanganate. The oxidizer and fuel tanks were pressurized by compressed air and compressed nitrogen, respectively, stored in high-strength cylinders. Four adjustable graphite control vanes were used; one vane was placed in the exhaust of each engine chamber. The main body of the missile was constructed from a lightweight, high-strength aluminum alloy.

The fuel tanks were made from nonreinforced cylindrical shells mounted between two bottom plates that were segments of spheres. The oxidizer tank was mounted forward of the fuel tank to control the center of gravity during flight. Moreover, an intermediate bottom plate was also mounted in the oxidizer tank for the same purpose, so that additional oxidizer would flow from the top to the bottom portion of the tank as the oxidizer was consumed from the bottom portion. The oxidizer feed line ran through a tunnel pipe built into the fuel tank. A riveted instrument compartment was located between the two tanks (Fig. 5). To shift the aerodynamic center of force closer to the center of mass, four fixed aerodynamic stabilizers were placed on the tail section of the missile.

This missile was flight-tested at the Kapustin Yar facility from July 1957 to December 1958. During this time, there were 24 launches of this missile, which confirmed that it was highly reliable. Some final design modifications were made, and the 8K63 missile was accepted into armaments for use with aboveground and silo launchers. The silo-launched version of this missile was assigned the code number 8K63U.

The requirement to develop both aboveground and silo-based launch facilities for this missile was largely dictated by the need to reduce the amount of time required for final development of the missile. Constructing a silo launcher would have required a large amount of time, so most test launches of this missile were done from a hastily constructed aboveground launch facility.

The tactical, engineering, and operational characteristics of the 8K63 represented a considerable advance over previous missiles. The 8K63 had a reaction time of 20 min and could deliver a 2.3-MT warhead to a maximum range of 2080 km. For some time, this was the main missile used by the Strategic Missile Forces (established December 1959). At the same time the 8K63 was on

operational duty in the Strategic Missile Forces, it also served for 25 years as the primary launch vehicle for testing new technology and designs of warheads and antimissile defense systems. The 8K63 was decommissioned in July 1988 pursuant to the Treaty on the Elimination of Intermediate-Range and Shorter-Range Missiles (INF Treaty).

11K63 (SL-7) Launch Vehicle (Fig. 6). Even the first few launches of spacecraft into low Earth orbit had demonstrated a wide variety of new opportunities to study Earth and near-Earth space using space-based instrumentation. These opportunities stimulated wide interest among scientists, businesspeople, and the military in obtaining information on Earth's surface, Earth's upper atmosphere, Earth's magnetic field and cosmic rays, the interaction between these particles and Earth's magnetic field, and the effects of space environment on objects launched into space.

This generated a requirement to launch a large number of spacecraft into low Earth orbit for various purposes. A need for low-cost launch vehicles therefore arose. The three-stage launch vehicle then (late 1950s) in use in the Soviet Union, the 8K72 Vostok, was not appropriate for frequent use as a launch vehicle due to the relatively high cost of launch, the relatively large amount of time required to prepare it for launch, and the fact that it was generally used to address more prestigious problems. Thus, in late 1959, OKB-586 embarked on an initiative to develop a two-stage launch vehicle based on the mass-produced 8K63 missile. This proposal was supported by the USSR Academy of Sciences and Ministry of Armaments, each of which were interested in the development of an inexpensive launch vehicle to address their specific needs using small spacecraft placed in low Earth orbit.

OKB-586 received the Government order to develop this launch vehicle, which came to be called the 11K63, in August 1960. The Government authorized the production of ten 11K63 launch vehicles and use of these vehicles to launch 10 spacecraft; each was for a different purpose and carried different instrumentation. Two of these spacecraft, designated the MS series, were developed by OKB-1, and the remaining eight spacecraft, designated the DS series, were developed by OKB-586.

The main tasks required for developing the 11K63 launch vehicle involved developing a second stage and aerodynamic fairing and adapting these components to a first stage that was virtually identical to the 8K63 missile. The second stage and aerodynamic fairing weighed ~ 7.7 metric tons, giving the 11K63 launch vehicle a launch weight of 49.4 metric tons, and a length of 30 m.

The second-stage fuel and dry compartments of the launch vehicle were similar to the corresponding first-stage compartments. However, there were also several differences due to the fact that the second-stage RD-119 engine required liquid oxygen and unsymmetrical dimethylhydrazine (UDMH) as propellants. This engine had been developed by OKB-456 for use on the Vostok launch vehicle but ended up not being used for a variety of reasons. The RD-119 engine was fairly well developed and also had relatively good energy performance characteristics (thrust and specific impulse in vacuum 106 kN and 3454 N·s/kg, respectively). The existence of these additional propellants undoubtedly made operation of the launch vehicle more complicated, but the availability of a fully developed engine reduced the effort and time required to develop the launch

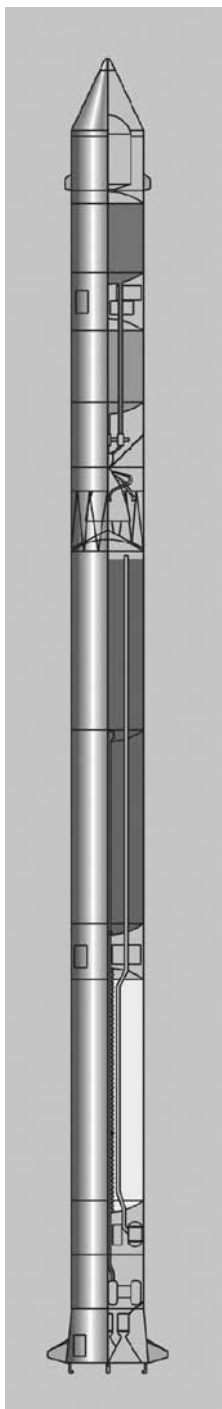


Figure 6. 11K63 launch vehicle. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

vehicle. Therefore, OKB-586 decided to use the RD-119 engine in the second stage of its first launch vehicle.

The RD-119 engine is a fixed, single-chamber, liquid-fueled engine installed on the second stage in combination with several movable low-thrust nozzles used to control the second stage in pitch, yaw, and roll. The engine is started using a pyrotechnic device. Positive pressure in the oxidizer tank was maintained by evaporating oxygen in a heat exchanger mounted on the engine's turbine exhaust pipe. Positive pressure in the fuel tank was maintained by using a mixture of producer gas and UDMH vapor. The RD-119 engine was only capable of single-use operation; as a result, spacecraft were launched to place them directly into orbit—primarily low-level highly elliptical orbits. To increase the amount of UDMH that could be stored in the second-stage fuel tank, it was initially cooled to -45°C . The second stage was mated to the first stage using a tubular beam that had a conical heat shield attached to the lower chord to protect the first stage from the exhaust of the RD-119, as it pushed away the first stage during the stage separation process.

The spacecraft was initially housed under a conical/cylindrical aerodynamic fairing (jettisoned during the boost phase of the flight after passing through the dense layers of the atmosphere). The spacecraft was separated by using pusher springs. Like the first stage, the second stage of the launch vehicle had an autonomous onboard control system developed by the newly established OKB-692 in Kharkov (now NPO Khartron-Arkos) under the direction of Chief Designer B.M. Konoplev). Initially, the 11K63 was to be launched from the 8K63U launcher at the Kapustin Yar Test Site, and an appropriate operational scenario was developed for the launch vehicle under this assumption (including use of a silo launcher that was shorter than the launch vehicle).

The first launch of the 11K63 from a silo took place on 27 October 1961. Both the first and the second launches were unsuccessful. Nearly 5 additional months were required to eliminate all of the problems. The third launch of the 11K63 took place on 16 March 1962 and was successful. The first spacecraft designed and built by OKB-586 personnel, the DS-2, had been placed into Earth orbit. After 37 standard 11K63 silo launches from the Kapustin Yar Test Site, all further 11K63 launches were from a new, aboveground facility at the Plesetsk Test Site, which was developed by the Design Bureau for Transportation Machinery directed by Chief Designer V.N. Sobolev.

There were several differences between the operational configuration of the 11K63 for silo launches and that for surface launches. During a silo launch, final assembly of the launch vehicle occurred during placement in the silo. The first and second stages (including spacecraft) were tested in the launch support facility, transported to the launch facility separately in the horizontal position, and then placed in the silo in the correct order. During a surface launch, final assembly of the launch vehicle took place in the launch support facility, and the assembled launch vehicle was transported to the launch facility, where it was raised into a vertical position using special equipment rather than the gantry.

The 11K63 could place a payload of up to 450 kg into circular low Earth orbit (200 km altitude and inclination 82°). The 11K63 became the first Soviet launch vehicle to be mass-produced. It was accepted for operational use in 1965, along with the DS-P1-Yu spacecraft (developed by OKB-586) and the new

aboveground launch facility at the Plesetsk test facility. The 11K63 was launched a total of 165 times, of which 143 were successful. Numerous Kosmos- and Interkosmos-series spacecraft were launched using the 11K63. It was used for 16 years until the final launch on 18 June 1977.

8K65 (SS-5) Missile and 11K65 Launch Vehicle (Fig. 7). The 8K65 became the second missile developed by OKB-586 within its new conceptual framework. The government task order for developing this missile was received in early July 1958. The goal was to develop, within 2 years, a missile that had

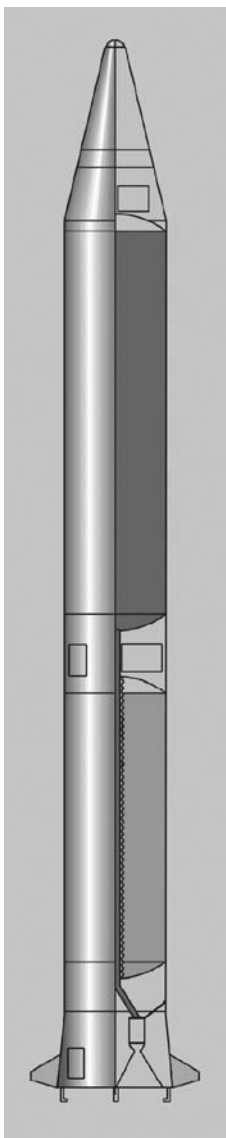


Figure 7. 8K65 missile. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

twice the range of the 8K63, which was then in the final stages of development. This promised to be a more difficult effort than the 8K63 had been and would be made even more difficult by the fact that OKB-586 had already been working for a year and a half on developing the 8K64 intercontinental ballistic missile, for which OKB-586 had received the task order in December 1956.

However, gaining experience, OKB-586 enthusiastically began work on developing the 8K65. The only decision valid at that point was made: Use the engineering design solutions that had proved themselves during the 8K63 design effort, as well as several of the basic design solutions that lay at the core of the 8K64 design. It was decided to use a monocoque, single-stage design for the 8K65 similar to that for the 8K63, but using a higher energy propellant combination. The 8K63 used the same propellant combination used in the 8K64; unsymmetrical dimethylhydrazine (UDMH) was the fuel, and AK-27I was the oxidizer. This propellant combination enabled a 15% increase in the total energy output of the 8K65 over the TM-185/AK-27I propellant combination. The use of UDMH enabled the entire missile to operate on only two propellants, thereby eliminating one deficiency of the 8K63—the fact that several propellants were required. One important quality of the propellants selected was that they would ignite on contact and thereby eliminated the need for an ignition propellant. This also simplified development of the turbopump-boost gas generators, which could use the same propellant combination.

The opportunity to use UDMH in the 8K65 resulted from the success achieved by the State Institute for Applied Chemistry (director V.S. Shpak) in researching this fuel and setting up the appropriate manufacturing facilities. However, the decisive factor enabling its use in the 8K65 was that the engine designers at OKB-456 had developed a two-chamber, liquid-fueled rocket engine (that had a thrust of 735 kN) that used this propellant combination. Two of these engines together as a single engine module [given the code number RD-216, whose thrust and specific impulse were 1481.3/1741.3 kN and 2413.3/2835.1 N·s/kg, respectively (sea level/vacuum)], were used on the 8K65.

The increased range required a corresponding increase in the amount of fuel and oxidizer carried, and thus, an increase in the diameter of the missile. To standardize production, a diameter of 2.4 m was adopted for the 8K65 main body; this was identical to the diameter of the 8K64 second stage.

The 8K65 was designed by virtually the same team of developers as the 8K63, and there were significant similarities to the 8K63 in both design and external appearance. Like the 8K63, the 8K65 was a single-stage monocoque design that had a conical warhead compartment (containing the same nuclear warhead as was used on the 8K63); a conical warhead mounting adapter, cylindrical fuel and oxidizer tanks; an instrument compartment containing the autonomous onboard control system that was located between the tanks; and a conical tail section that had a fixed RD-216 engine module. The 8K65 used exactly the same tank pressurization system, actuator-control design, and structural materials as the 8K63. Just as in the 8K63, the oxidizer tank was located forward of the fuel tank, and four fixed stabilizers were mounted on the tail section. The major differences between the 8K65 and the 8K63 were the new fuel; the increased fuel capacity; the larger diameter of the missile; the new, more powerful engine; and the new control system.

The 8K65 engines were ignited by taking advantage of the hypergolic nature of the propellants as they mixed in the combustion chamber, thereby simplifying the engine design. Using the same basic fuel/oxidizer components also enabled simplifying the gas generators used in the turbopump-boost system.

However, there were also some additional differences. For example, the cylindrical fuel tank fairings on the missile were made from reinforced molded panels. A system for simultaneously emptying the tanks (not available on the 8K63) was used to reduce the amount of unused fuel and oxidizer by synchronizing the consumption rates. This missile also marked the first use of a gyro-stabilized platform, that gave the missile the same accuracy as the 8K63, despite the factor of 2 greater range. The warhead unit was separated by braking the missile body using solid-rocket motors mounted on the missile instrument compartment housing.

The 8K65 was twice as heavy as the 8K63, but was only 2.3 m longer. It could deliver a warhead identical to that used in the 8K63 to a maximum range of 4500 km. Like the 8K63, the 8K65 was developed and placed into military service for use from surface and silo-based launch facilities. The silo-launched version had the code number 8K65U. Silo launches of the 8K65U were similar to those of the 8K63U; they involved the use of a launch canister and the gas pressure generated by the missile engines after being started in the silo. The 8K65 was flight-tested at the Kapustin Yar Fourth State Central Test Site. Forty-four launches of the 8K65 or 8K65U missile were performed as part of the flight/development testing program. The 8K65 became the second, longest range missile in the Strategic Missile Forces arsenal. The 8K65 enabled targeting strategic facilities maintained by potential adversaries in Europe, Asia, and even Africa that were out of range of the 8K63. The missile remained on operational duty for 15 years, starting in 1962, and was decommissioned in 1987 pursuant to the INF Treaty.

11K65 (SL-8) Launch Vehicle (Fig. 8). The second launch vehicle, designated the 11K65, was developed by OKB-586 pursuant to a USSR Government Decree issued on 30 October 1961. According to the stated requirements, the 11K65 launch vehicle would launch one (or more) spacecraft into Earth orbit. The orbits were required to be either elliptical or circular, the circular orbits had a maximum altitude of 2000 km. Stringent requirements were also imposed on the accuracy of the resulting orbital parameters.

The spacecraft to be launched by the 11K65 were expected to remain active for long periods of time, and therefore would weigh more. In view of all these requirements, the 11K65 had to have approximately three times the total spacecraft launch capacity of the 11K63. Preliminary studies performed by OKB-586 indicated that such a launch vehicle could be built quickly and at minimal expense in terms of materiel, if the 8K65 missile were used as a first stage. Design optimization of the launch-vehicle parameters revealed that it would be desirable to design the second stage so that it would have the same diameter, use the same propellants as the first stage, and weigh ~24 metric tons. Placing such a stage on the 8K65 missile required partially modifying its frame configuration and developing a new interstage compartment that could bear the load of the second stage and also provide an outlet for the gas generated by the steering nozzles on the second-stage engine during stage separation. To reduce the length and

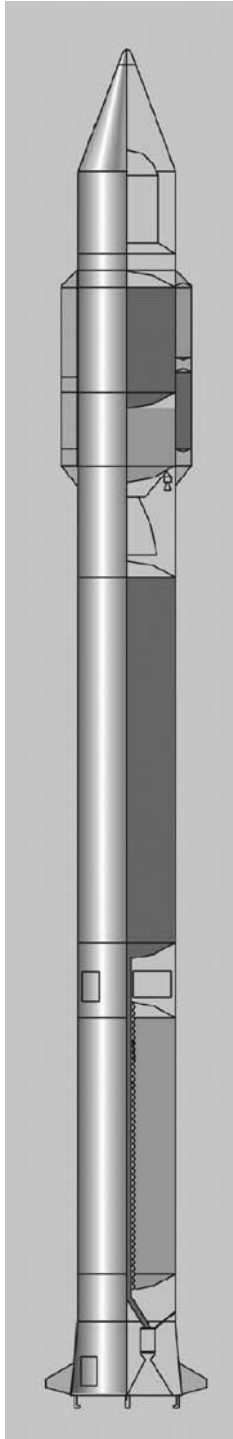


Figure 8. 11K65 launch vehicle. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

weight of the second stage, it was decided to use a somewhat different design compared with that of the first stage. The second-stage fuel tanks were combined into a single cylindrical propellant compartment that had three spherical-segment bottom plates and a common intermediate bottom plate separating the propellant compartment into an upper cavity containing the oxidizer and a lower cavity containing the fuel. The instrument compartment was located above the propellant compartment, and a spacecraft adapter and clamshell nose fairing were attached to the top of the instrument compartment to protect the spacecraft against the air flow while the launch vehicle moved through the lower atmosphere and to protect the spacecraft during ground operations.

The requirements that the 11K65 reach a high-altitude circular orbit and launch heavy spacecraft into such orbits led to the use of a two-burn orbital insertion procedure for placing spacecraft into such orbits. In this procedure, the launch vehicle second stage and spacecraft were placed into an elliptical transfer orbit whose apogee was equal to the desired altitude of the circular orbit, and were then transited into the desired circular orbit. The problem came in obtaining the appropriate magnitude and direction of the velocity vector required for the second stage to reach the perigee of the elliptical transfer orbit and then the circular orbit. Implementation of this procedure affected the design concepts for the control system, the second-stage engine, and the propellant feed system.

The second-stage engine, which came to be known as the S5.23, actually consisted of two subengines: a single-chamber main engine and a four-chamber steering engine. The main engine was fixed relative to the missile. The four steering-engine nozzles were combined with the four small thruster nozzles to form four combined, steerable units mounted on the main engine. The steering engine enabled control of the second stage while the main engine was in operation; the thruster system was used for stabilization and attitude control of the second stage during the passive portion of the trajectory. This engine was developed by OKB-2 (now known as the Chemical Machinery Design Bureau) under the direction of Chief Designer A.M. Isaev. Features of the S5.23 engine included three thrust modes (vernier, intermediate, and primary) and the capability of reusing the primary-thrust mode. In primary mode, the engine thrust was produced by the main engine and its four steering nozzles and was equal to 157 kN (specific impulse 2913.6 N·s/kg). In intermediate-thrust mode, only the steering nozzles were used (thrust 5.4 kN). This mode was used during engine start-up and shutdown. The low-thrust mode (thrust 0.1 kN) was supported by the passage of gas produced in a special gas generator through small nozzles. Between the two engine firings, the fuel and oxidizer for the second engine firing, and for the engine firing in the stabilization/attitude control mode during the passive portion of the trajectory, was stored in two small tanks located on the outside of the second-stage propellant compartment.

The oxidizer and fuel tanks were pressurized by using compressed air and compressed nitrogen, respectively, stored in high-pressure cylinders.

The control system for the second stage was developed by OKB-692 under the direction of Chief Designer V.G. Sergeev. To meet the requirements for increased accuracy of spacecraft placement into orbit, relatively high-speed computers and high-accuracy control devices (that had increased throughput and special software and algorithms) were used for the first time. The engineering

design solutions implemented for the 11K65 gave it the ability to place up to eight spacecraft that had a maximum total weight of 1500 kg into Earth orbit (circular orbit, altitude 200 km, and inclination 51°) in a single flight. OKB-586's role in developing the 11K65 was limited to preparing the preliminary design and design documentation, as well as fabrication and design/flight-testing of the first 10 launch vehicles. Due to the workload related to developing the new 8K67 missile, the 11K65 development task was transferred to OKB-10 (now known as the Academician M.F. Reshetnev Scientific Production Association for Applied Mechanics) in late 1962.

Development flight-testing of the 11K65 began on 18 August 1964 at the Baikonur Cosmodrome from a modified aboveground launch facility designed by the Novokramatorsk Machinery Plant Design Office. Routine 11K65 operations have been conducted at the Plesetsk Test Site since 1967 and the Kapustin Yar Test Site since 1973. Aboveground launch facilities that had movable gantries were built at each of these test sites. These launch facilities were designed by the Design Bureau for Transportation Machinery. Between 1965 and 1967, OKB-10 upgraded the 11K65 to improve its operational specifications, after which it was renamed the 11K65 M (Kosmos-3). To date, there have been more than 500 launches of the 11K65 or 11K65 M that carried more than 1000 Kosmos- and Interkosmos-series spacecraft into Earth orbit, including more than 130 DS-1P, Tselina, Tyul'pan, Taifun, and AUOS ("automated universal orbital stations") spacecraft designed by OKB-586 (Yuzhnoye Design Office). The 11K65 remains one of the most reliable, highly productive launch vehicles in the Russian Federation inventory. For some time, a modified version of this launch vehicle, the K65M-R, has also been used for testing warhead units and missile-defense countermeasures.

8K69 (SS-9 mod 3) Missile—The Basis for the 11K69 and 11K68 Launch Vehicles (Fig. 9). The Government task order for developing the 8K69 missile was issued on 14 April 1962, simultaneously with the task order for developing the 8K67. The 8K67 was given an accelerated development schedule. Many of the requirements for the 8K67 and 8K69 missiles were either very similar or identical, so the decision was made to standardize these missiles in areas where the engineering design solutions were consistent. Thus, both stages of the 8K67 were used in the 8K69, after a few slight design modifications, mainly related to the fact that the 8K69 used a warhead different from the 8K67.

The 8K69 was developed as a countermeasure to U.S. deployment of the Safeguard missile defense system, which protected U.S. territory against missile attack from the north. This missile had several unique features. It had unlimited range and could reach targets from any direction; therefore, it could deliver a nuclear warhead to any area of the United States by evading the missile defense system (and rendering it useless). At the time, these features led to calling the 8K69 a "global missile." The 8K69 warhead could come in from various directions by being placed into Earth orbit at various azimuths and then reentering from orbit and hitting the target.

Like the 8K67, the 8K69 was a two-stage monocoque design in which the stages were on top of each other and stage separation occurred around the circumference of the missile (launch weight 181.3 metric tons and length 32.65 m). Both missile cases were cylinders 3 m in diameter. Both stages (and the warhead)

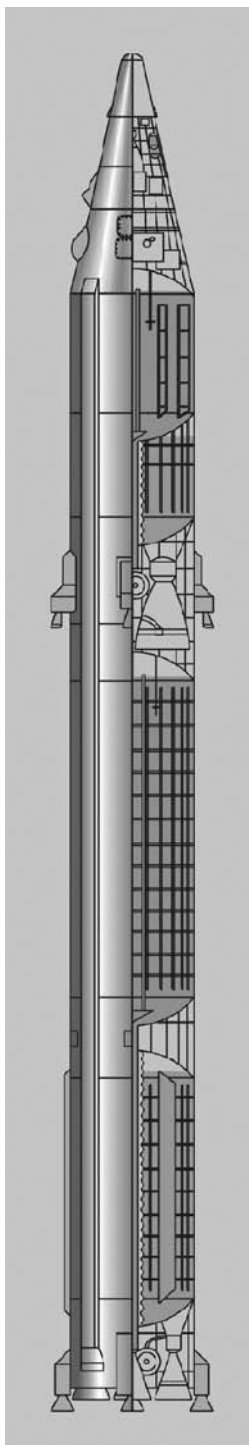


Figure 9. 8K69 missile. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

used the same fuel/oxidizer pair: unsymmetrical dimethylhydrazine (fuel) and nitrogen tetroxide (oxidizer). This fuel/oxidizer pair gave the missile a high total energy output and also substantially increased the guaranteed military readiness lifetime (to 7 years).

The first stage consisted of four sections: a stage transition section, an instrument section, a tail section, and propellant-tank section. The second stage had three sections: an instrument section, a propellant section, and a tail section. The second-stage propellant tanks were combined into a single section that had three spherical-segment bottom pieces. The second-stage instrument section was conical. All “dry” sections (stage transition section, instrument sections, and tail sections) in the missile and the warhead section were riveted, whereas the fuel tanks were welded. The fuel-tank design used molded panels and molded stock hollow frames, which led to a substantial reduction in the number of processes required, as well as an overall simplification of the tank fabrication process. To reduce the structural weight, light, high-strength aluminum alloy was used for all missile and warhead sections.

The first-stage propulsion system consisted of an RD-251 main engine and an RD-855 steering engine (total thrust 2651.6/2974.3 kN and specific impulse 2627.1/2945.9 N·s/kg at sea level and in vacuum, respectively). Structurally, the RD-251 main engine consisted of three identical engines affixed to a single frame. Each of these engines had two chambers, a turbopump unit, a gas generator, a solid-fuel starter, automatic control systems, and various other components. The RD-251 engine was attached to a frame, which was in turn attached to the fuel tank bottom frame, and was enclosed by the tail section. Solid-fuel starters were used to start all three engines simultaneously within a few seconds of steering engine start-up. The engines could also be turned off simultaneously a few seconds before the steering engine, in response to a command from the command system.

The RD-855 steering engine controlled the missile in roll, pitch, and yaw. It consisted of four steerable chambers and a stationary turbopump unit, gas generator, solid fuel starter, automatic control systems, and various other components. The steering engine chambers were steered by using hydraulic actuators: UDMH was obtained from the motor turbopump unit as the working fluid. The steering motor was started and turned off by the missile control system in accordance with a timed cycle schedule. The steering engine chambers were mounted in four fairings on the exterior surface of the tail section. Two of these fairings also contained solid-rocket motors to decelerate the first stage after stage separation.

The second-stage propulsion unit consisted of an RD-252 main engine and an RD-856 steering engine (total thrust 9955.7 kN and specific impulse 3093.1 N·s/kg in vacuum). The RD-252 design and systems were similar to those used in the first-stage, two-chamber, liquid-fuel engine but had larger nozzles for a higher maximum altitude. Like the RD-251, the RD-252 was affixed to an engine frame that was in turn attached to the fuel tank bottom frame, and was enclosed by the second-stage tail section.

The RD-856 steering engine was designed to control the second stage, which operated in a less-dense atmosphere. Therefore, the engine had much lower thrust than the RD-855 but was similar in design and was based on similar

systems. The second-stage engine was mounted similarly to the first-stage engine. The steering chambers were likewise housed in four compartments mounted on the exterior surface of the second stage. Solid-propellant motors were also mounted in two of the compartments; these motors were used for second-stage braking after warhead separation. Second-stage engine start-up and operation were supported by using a control system similar in design to that used for first-stage engine start-up and operation. The main engines for both stages were developed by OKB-456, the steering engines were developed by OKB-586, and the retroengines were designed by the Iskra Machinery Plant in Moscow.

The fuel and oxidizer tanks for each stage were pressurized by using hot gases produced from the primary propellants using special gas generators (the steering-engine propellant feed system was a source).

The 8K69 autonomous onboard control system supported prelaunch missile preparation and silo-based launches. The control system also supported fairly high readiness and target accuracy. To improve the reaction time of the missile, the control system used gyro units that could be forced into operational mode. The orbital weapon unit (OWU) consisted of a relatively large single warhead, as well as a reentry stage and an interstage bay that connected the orbiting weapon unit to the second-stage instrument bay. The OWU reentry stage included an instrument bay that held the control system and toroidal fuel tanks; an RD-854 retro-engine and vernier thruster were located in the interior cavity between the tanks. The RD-854 is a fixed, single-chamber liquid-fueled engine (thrust 75.5 kN, specific impulses $3063 \text{ N} \cdot \text{s/kg}$), that uses eight fixed exhaust nozzles to control the orbital weapon unit during braking maneuvers. This engine was started by using a pyrotechnic starter. The vernier thruster (8 nozzles, thrust 0.03 kN) was used for OWU damping after OWU separation from the second stage, for OWU trajectory stabilization during passive flight, and for OWU attitude stabilization before retroengine start-up. The RD-854 engine and thruster were designed by OKB-586. The control systems for the missile and orbital weapon were designed by OKB-692, and the gyro instruments were developed by NII-944.

The missile was silo-launched using the gas pressure generated by the first-stage missile engines after they were started in the silo. The missile was launched from a stationary launch platform located inside the launch silo. After the missile left the silo, it was turned toward the target azimuth in roll using the first-stage steering engines. Missile-mounted hooks slid along guides in the launch canister to assure the safety of missile launches. In the interests of launch safety while the engines were in operation, the gas outflow from the first-stage engines was directed away from the missile by special baffles.

Ground-based firing-stand tests of the OWU reentry stage and aircraft-based testing of the OWU reentry stage under weightless conditions were followed by development flight-testing of the 8K69 missile, which involved some 19 missile launches. Final design modifications of the first and second stages had already been made as part of the 8K67 missile development project, so flight-testing of the 8K69 was limited in scope and basically involved integration testing of the OWU and missile. The 8K69 was accepted into armaments in 1968 and remained in military service for 15 years.

11K69 (SL-11) and 11K68 (SL-14) Launch Vehicles (Figs. 10, 11). OKB-586 began developing the 11K68 and 11K69 launch vehicles in August 1965

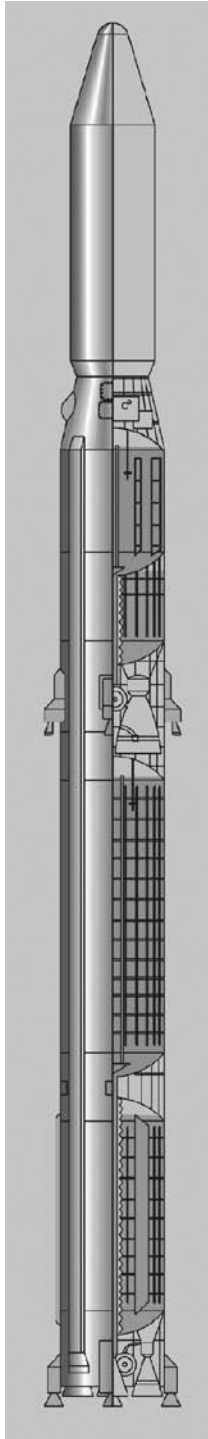


Figure 10. 11K69 launch vehicle. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

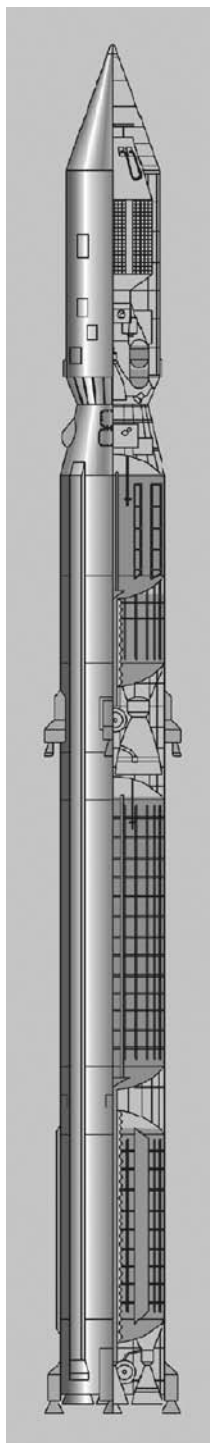


Figure 11. 11K68 launch vehicle. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

pursuant to a government decree. The USSR Ministry of Defense, the main party interested in developing these launch vehicles (LVs), hoped to use them as delivery systems for both long-term or tactical space-based reconnaissance systems and antisatellite defense systems. As a result, these launch vehicles had to be extremely flexible and rapidly launched. The 8K69 missile was selected as the basis for both LVs. Virtually identical versions of the first two stages of this missile were used in both the 11K68 and 11K69 LVs. This led to standardization of many launch-support-facility and launch-facility components for these LVs, as well as a reduction in development cost.

There were various reasons behind selecting the 8K69 missile as the basis for the 11K68 and 11K69 launch vehicles. The 8K69 missile, which could place orbital warheads into Earth orbit, could already, after some modification, place other spacecraft into orbit. Additionally, the 8K69 had a relatively high total energy output and could place payloads of up to 3000 kg into low Earth orbit. It could also be readied for launch fairly rapidly, thereby laying the groundwork for meeting the 11K68 and 11K69 requirements for flexibility and rapidity of launch.

Another consideration was the fact that the first and second stages of the 8K69 were already in final development as part of the 8K67 missile and the necessary series production facilities had been established for them. This enabled reducing the one-time development costs for the 11K68 and 11K69 LVs. These considerations served as the basis for the decision to base the 11K68 and 11K69 LVs (which later came to be called, respectively, the Tsiklon-3 and Tsiklon-2) on the 8K69 missile.

The 11K69 LV is an 8K69 missile in which the orbital warhead has been replaced by one of two nose units containing either a reconnaissance spacecraft or an anti-satellite system. The anti-satellite system consisted of a guidance stage [developed by the Central Machinery Design Bureau (TsKBM), Chief Designer V.N. Chelomei], and an antisatellite weapons system (developed by Comet Design Bureau, Chief Designer A.I. Savin). Each of these standard 11K69 payload types can be placed under a standard nose fairing.

The launch vehicle is assembled in the Launch-Support-Facility Test and Assembly Building, with the launch vehicle and nose units in horizontal position. The 11K69 LV is transported to the launch pad in assembled form using a custom transporter-erector. Various assemblies installed on the transporter-erector ensure proper mating (or demating) of all pneumatic lines, hydraulic lines, electrical lines, and mechanical joints, the appropriate lines and connections to the launcher are made as the transporter-erector moves over them. The LV is placed into vertical position using the transporter-erector, which is also used for all further servicing (no gantry tower is used).

The 11K69 LV marked the first use of a safe, automated, crewless launch system that completely eliminates any need for operations personnel in the vicinity of the launch facility during the most hazardous launch operations. This system of operations also led to a substantial reduction in the amount of time required to prepare the LV for launch.

The 11K69 space launch system was developed in cooperation with the Design Bureau for Transportation Machinery, the Elektropribor Design Bureau, and various other organizations. The space launch system for the 11K69 LV was located at the Baikonur Cosmodrome (Scientific Research Test Site 5). Standard

11K69 launches have been performed since 6 August 1969. This LV holds a unique record in rocketry and space history. All 103 launches of this launch vehicle have been successful, and the 11K69 is still in regular use.

The 11K68 LV differs from the 11K69 LV in that the former has a newly developed third-stage, nose fairing, and interstage-skirt section. This interstage-skirt section consists of an inverted frustrum of a cone because the mating diameter of the nose cone was greater than that of the second-stage instrument bay. The nose cone is intended to protect the spacecraft and third stage from external effects during ground and flight operations. The nose cone consists of a conical cylindrical “clamshell” whose two halves are held together by a special locking device embedded in the longitudinal surface of the nose cone. Once the dense layers of the atmosphere have been traversed, the mechanical linkage between the two halves is severed, and they are jettisoned from the launch vehicle.

The third stage of the 11K68 launch vehicle includes a spacecraft adapter, an instrument section, fuel/tail section, RD-861 main engine, 11D75 vernier thruster, a control system, and a telemetry system. To reduce the weight of the third stage, the instrument compartment uses a space-frame design. In addition to the control system devices, the instrument compartment also houses a spherical high-pressure tank containing helium used by the propellant-tank pressurization system. The telemetry system instruments are mounted on the exterior of the instrument section, along with the all-riveted spacecraft adapter structure. The third stage uses a number of engineering design solutions tested during basic design of the orbital-weapon-unit (OWU) braking engine assembly on the 8K69 missile. Like the OWU, the third stage was sealed. The third stage also used the same propellants as the OWU, nitrogen tetroxide and unsymmetrical dimethylhydrazine. The third stage can be stored for several years in the fueled and gas-filled state and transported to the launch pad in this state as part of a launch vehicle. This engineering design solution led to a significant reduction in launch preparation time. Both the third stage and the OWU used a toroidal fuel compartment that had a fixed main engine mounted in the interior cavity. The thruster was also fixed and was mounted in the third-stage tail compartment.

By contrast with the OWU, the third stage toroidal fuel section was cylindrical rather than conical and had one and a half times the volume. A common intermediate bottom plate divides the section into an upper cavity for the oxidizer and a lower cavity for the fuel. The cavities include baffles to prevent propellant oscillation, intake assemblies, and other fittings. Fine-mesh mist extractors were used to support propellant feed into the main engine during start-up under weightless conditions.

The RD-861 main engine is an improved version of the RD-854 engine used in the OWU and has a higher total energy output (thrust 78.1 kN, specific impulse 3374 N·s/kg) than the RD-854. However, its main features are its ability to be fired twice under weightless conditions and extended operational life. These features enabled using a two-impulse spacecraft launch trajectory that had an elliptical transfer orbit and resulted in the ability to increase the altitude of the orbits that could be reached, as well as the mass of the spacecraft that could be placed in such orbits.

The RD-861 main engine is a single-chamber, liquid-fueled engine that has an open-loop turbopump-based propellant feed system; the generator gas used to

operate the turbopumps is discharged through eight nozzles operated by the third-stage flight-control actuator system while the main engine is in operation.

The 11D75 vernier thruster is an improved version of the OWU thruster on the 8K69. Just as for the main engine, both the number of times the thruster can be operated and the operating lifetime have been improved; the functions performed by the thruster were also expanded. It was used for attitude control and stabilization of the third stage and also to support main-engine restart in weightless conditions. To this end, the thruster was equipped with two special nozzles to create a micro-overload in the longitudinal direction for approximately 100 seconds until the third-stage main engine restarted. The 11D75 thruster has an independent fuel system that is filled from the primary third-stage fuel tanks while the first two stages of the 11K68 launch vehicle are in flight. The RD-861 and 11D75 engines were developed by the Yuzhnoye Design Office.

The 11K68 control system consists of two interconnected systems; one has equipment installed in the first- and second-stage instrument sections, and one has equipment installed in the third-stage instrument sections. The control system for the first two stages of the launch vehicle supports prelaunch preparations, launch, and flight-control functions until third-stage separation, and the second control system controls the third stage during subsequent spacecraft orbital insertion phases. The control system for the first and second stages of the LV was developed by the Elektropribor Design Bureau; that for the third stage was developed by the Kiev Radio Plant (now the Kiev Radio Plant Production Association) Design Bureau. Like the 11K69, the 11K68 was operated in horizontal mode. Like the 11K69, preparation of the 11K68 for launch was automated using standardized assemblies, command lines, actuator lines, and data lines.

The ground facility for these launch vehicles was developed by the Design Bureau for Transportation Machinery and located at the Plesetsk Test Site. To date, there have been 119 launches of the 11K68 launch vehicle; 114 were successful. Several launches have supported the deployment of up to six spacecraft in one flight. More than 10 different types of spacecraft have been deployed into a variety of orbits using the 11K68.

15A18 (SS-18 mod 4) Missile (Fig. 12). The 15A18 is one of the most powerful ballistic missiles ever developed by the Yuzhnoye Design Office; it is designed to deliver 10 warheads weighing approximately 8.5 metric tons to a range of up to 11,000 km. The missile had a launch weight of 211.1 metric tons, a length of 34.3 m, and a case diameter of 3 m. One such missile can destroy up to 10 arbitrarily located targets within an area several hundred kilometers across using the nuclear warheads carried by the missile. The 15A18 has the capability of high-accuracy strikes on enemy facilities even after multiple enemy nuclear strikes against 15A18 deployment areas.

The government task order for developing the 15A18 missile system was issued on 16 August 1976. The new missile was to be based on the 15A14 missile system, and the main reason for developing the new system was to modernize the 15A14 for increased military effectiveness and increased launch-facility hardening. The improved 15A14 missile was designated the 15A18. The task order called for increasing the yield of the nuclear warheads, improving the

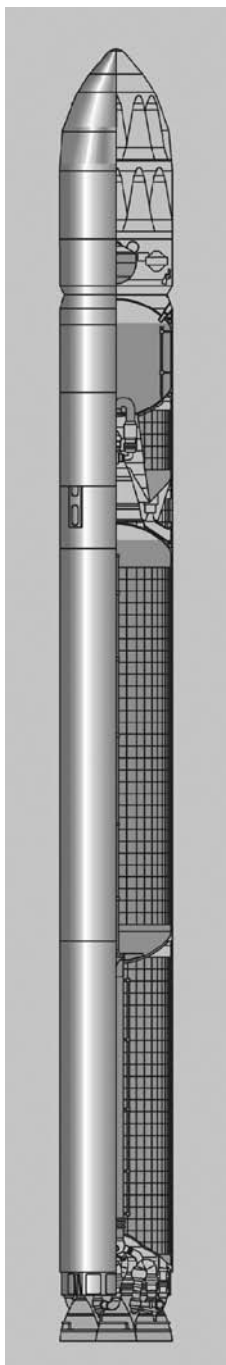


Figure 12. 15A18 missile. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

effectiveness of the missile-defense countermeasures used, increasing the target accuracy, increasing the range, and increasing the dimensions of the area targetable by the reentry vehicles to meet certain specifications. This task was successfully accomplished.

These requirements were met through the following developments:

- high-strength reentry vehicles carrying higher-yield nuclear warheads;
- multipurpose missile-defense countermeasures;
- two-level weapons compartment that has clamshell fairing;
- high-energy output liquid-fueled reentry-vehicle stage separation units and missile-defense countermeasures;
- missile control systems that have gyroscopes of enhanced accuracy and an improved onboard digital computer; and
- targeting systems that remain functional and ensure highly accurate targeting, even in the event of multiple enemy nuclear attacks on the 15A18 deployment area.

A launch silo that had improved hardening against nuclear explosions was also developed.

The 15A14 missile was used for the first and second stages of the 15A18 after several modifications. These missile stages have a highly compact layout to maximize the amount of propellant carried within the limited dimensions of the missile. The first-stage propulsion unit is a module consisting of four single-chamber RD-264 engines developed by the Design Bureau for Power Machinery (total thrust 4167.3/4524.4 kN and specific impulse 2877.3/3123.5 N·s/kg at sea level and in vacuum, respectively). The second stage uses an RD-229 single-chamber, fixed, main engine and an RD-0230 four-chamber steering engine (total thrust 760.3 kN and specific impulse 3193.5 N·s/kg in vacuum). Both of the second-stage engines were developed by the Design Bureau for Automated Chemical Equipment. To increase the specific impulse, the RD-264 and RD-0229 engines had a closed-loop design in which the gas-generator output used to operate the turbopump assemblies was also burned in the combustion chambers.

The RD-864 engine on the RV bus was developed by the Yuzhnoye Design Office (thrust 19.6 kN and specific impulse 2848 N·s/kg). It is a four-chamber, liquid-fueled engine built to a new design; before engine start-up, the chambers are moved outward from the body of the bus and are stopped at a certain specific angle to the longitudinal axis of the bus so that the stage appears to be drawn forward. This simplified the system for separating the warhead unit and missile-defense countermeasures system components from the RV bus.

The first stage is controlled during flight by gimbaling the RD-264 engine, and the second stage and reentry stage are controlled by gimbaling the RD-0230 and RD-864 steering chambers. The chambers are gimbaled using hydraulic actuators where UDMH is the working fluid (the UDMH is fed by the turbopumps for each of these respective engines).

The tanks were prepressurized before engine start-up using a chemical pressurization system based on injecting a metered amount of oxidizer into the fuel tank and fuel into the oxidizer tank. In-flight pressurization of the first two stages of the missile used hot gas produced in special gas generators. The reentry stage used a gas-cylinder-based pressurization system.

The missile stages and warhead were separated by using braking systems on the first and second stages based on the impulse generated by blow off of pressure from the fuel tanks in each missile stage.

The missile was launched mortar-style from a launch canister placed in the silo. Launching the missile mortar-style provided the following advantages: substantial simplification of both the shock-absorption system for the surface test and launch equipment and the launch-silo design; improved protection against nuclear blast effects; reduction in launcher cost; and a reduction in the time required for launch silo construction and for placing the missiles into military service. The missile was flight tested at the Baikonur Cosmodrome. The flight-testing program for the 15A18 included 19 launches, of which 17 were successful. **Dnepr Launch Vehicle (Fig. 13).** The SALT-I and SALT-II treaties authorized the use of decommissioned strategic missiles for spacecraft launches. In this context, the Yuzhnoye State Design Office proposed a design for a missile/space system, called the Dnepr, based on the 15A18 (RS-20B) missile, which was then being decommissioned. The project was supported by the National Space Agency of Ukraine and the Russian Space Agency.

An international corporation, Kosmotras, including the following Ukrainian and Russian enterprises, was established to pursue this project: the Yuzhnoye State Design Office, the State Enterprise Production Association Yuzhnyi Machine-Building Plant, the Khartron Joint-Stock Corporation, and the Design Bureau for Special Machinery. The resulting company took responsibility for developing, operating, and marketing the system. The project was based on the availability of a significant base of materiel that could be used to address the purpose at hand: more than 150 15A18 missiles on duty in launch silos and stored in arsenals, as well as the experimental 15A18 ground facility at Baikonur Cosmodrome. Implementation of this project required some slight modifications to the missiles, including the onboard control systems and the launch support facility/launch facility at Baikonur Cosmodrome.

The Dnepr launch vehicle is essentially a modified 15A18 in which the weapon section is replaced by a spacecraft payload. The high energy output and high orbital inclinations reachable by using the Dnepr launch vehicle will enable it to be used for placing communications, remote-sensing, and scientific satellites into low Earth orbit. The first launch of the Dnepr took place on 21 April 1999, placing the British UoSAT-12 spacecraft into orbit. The second launch of the Dnepr launch vehicle took place on 26 September 2000, placing the following five spacecraft into orbit: UniSat and MegSat-1 (Italy), SaudiSat-1A and SaudiSat-1B (Saudi Arabia), and TiunSat-1 (Malaysia).

Treating space research as a highly valuable engine of scientific, technical, and economic progress, the Yuzhnoye Design Office, which began production of rocketry and space hardware more than 45 years ago, is continuing its space activities in developing new spacecraft launch systems for the Ukrainian national space program and other international space programs.

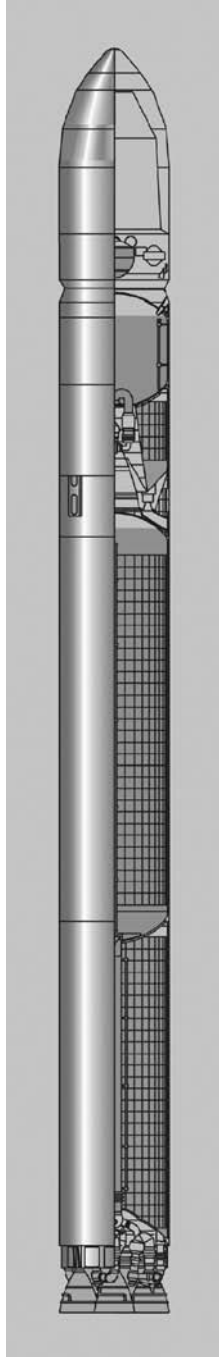


Figure 13. Dnepr launch vehicle. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

BIBLIOGRAPHY

1. *Yuzhnoye Design Office Missiles and Spacecraft*, Izd. OOO. KolorGraf/OOO RA Tandem, Dniepropetrovsk, 2001.

STANISLAV NIKOLAEVICH KONYUKHOV
M.K. Yangel' Yuzhnoye State Design Office
Dniepropetrovsk, Ukraine

COSMONAUTS SELECTION AND PREPARATION

Cosmonaut Training Center History and Accomplishments

The Soviet Union's launch of the world's first artificial Earth satellite on 4 October 1957 and its subsequent achievements in developing rockets and space technology and flying living beings in its space flight program marked the beginning of preparations for the first human spaceflight. The issue of human spaceflight was discussed in the USSR Academy of Sciences as early as the very beginning of 1959. The USSR Council of Ministers Resolution "On Preparing Humans for Space Flight" was issued on 22 May 1959.

In late 1959, at the initiative of S.P. Korolev, Chief of Rocket and Space Technology Design, a resolution was passed to create a special Center within the Air Force to train humans for the first piloted spaceflight. This Center was given the world famous name of Cosmonaut Training Center (further referred to as the Center). The founding date of the Center is traditionally considered 11 January 1960, when, in accordance with the previously mentioned resolution, the Air Force Commander in Chief published a directive that defined the organizational and personnel structure of the Air Force Cosmonaut Training Center. The site selected for this institution was a picturesque region on the outskirts of Moscow, currently known as Zvezdnyy Gorodok (Star City).

As the challenges confronting Soviet cosmonautics increased in complexity, the Center's structure and status altered. In 1965, the Air Force Cosmonaut Training Center was renamed the First Cosmonaut Training Center, and in 1968, it was named for the planet's first cosmonaut, Yu.A. Gagarin. In 1969, it was restructured as the Yu.A. Gagarin Scientific Research and Test Center for Cosmonaut Training and was granted the rights and status of a category 1 scientific research institute. In 1995, by Russian Federation (RF) Government Resolution, the former Gagarin Training Center and the 70th Individual Testing and Training Special Purpose Aviation Regiment were merged to create the Gagarin Russian State Scientific Research And Testing Center for Cosmonaut Training under the jurisdiction of the Russian Ministry of Defense (Air Force) and the Russian Aviation and Space Agency. The objective of this merger was to increase the efficiency with which the RF's scientific and technological potential in the area of piloted space flight and cosmonaut training was used to implement federal space programs and meet Russia's international commitments (Fig. 1).



Figure 1. The Center (bird's eye view). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Starting with Yu.A. Gagarin, the first person on Earth to fly in space, all Soviet, and then Russian cosmonauts, and also many cosmonauts and astronauts from foreign countries were trained and certified for space travel at the Cosmonaut Training Center. The most significant numerical indicators of the Center's performance are thus the flights and training of cosmonauts. Between 1961 and 2000, 90 crews and 125 cosmonauts flew on Russian spacecraft and space stations. Of these, 71 cosmonauts flew once, 27 twice, 21 three times, 3 four times, and 3 five times. Thirty-three flights included citizens of 18 foreign countries (Austria, Afghanistan, Bulgaria, Cuba, Czechoslovakia, France, Germany, Hungary, India, Japan, Mongolia, Poland, Rumania, Slovakia, Syria, the United Kingdom, the United States, and Vietnam.) The Center trained a total of 326 prime, backup, and reserve crews (a total of 447 cosmonauts of whom 78 were foreigners) for the flight programs of Russian spacecraft and stations.

Each year, on average, 24 cosmonaut candidates went through general space training, 22 cosmonauts underwent group training, and 12 crews (prime, backup, and reserve) underwent crew training at the Center. Cosmonauts were trained for the flight programs of the Vostok, Voskhod, and Soyuz spacecraft and the Salyut, Almaz, and Mir long-term orbital stations, the lunar program, and programs for testing the Energia-Buran space shuttle system and advanced aerospace systems. At the present time, crews are being trained for the International Space Station program.

Selection of the first cosmonauts began in October 1959 among men in Air Force units. It was believed that the best candidates for the first piloted space-flight would come from the ranks of military pilots. Documents for more than 3000 fighter pilots, aged 35 or below, were examined to select the first cosmonaut candidates. On the basis of interviews and outpatient medical examination, 206 pilots were allowed to move on to the medical selection phase. These candidates then underwent a final inpatient examination between October 1959 and April

1960. Only 29 pilots passed all phases of medical selection and, of these, 20 were accepted into the spaceflight training program (subsequently, requirements for cosmonaut health standards were relaxed somewhat).

The first group of cosmonauts trained at the Center included Yu.A. Gagarin, who completed an orbital flight of 1 hour 48 minutes on the Vostok-1 piloted spacecraft on 12 April 1961. After Gagarin's flight, G.S. Titov was trained to perform a 1-day space flight. This flight was completed in August 1961.

The next phase of the Center's activity focused on training cosmonauts for group flights. In August 1962, A.G. Nikolayev and P.R. Popovich completed the first such flight on Vostok-3 and Vostok-4. The program of training cosmonauts for flights on the Vostok series spacecraft was completed with the flights of B.F. Bykovskiy on Vostok-5 and the first female cosmonaut, V.V. Tereshkova, on Vostok-6 between 16 and 19 June 1963. The completion of the Vostok program provided a unique experience in training of cosmonauts for piloted flights.

Subsequently, the Center entered a phase in which cosmonauts were trained for flights on the multiman piloted spacecraft of the Voskhod series, which had just been developed. Training these crews emphasized the use of the new capabilities these spacecraft had for conducting scientific, technological, and biomedical experiments and enabling extravehicular activity. Only two flights were made within the Voskhod program. The first flight, on Voskhod-1, was piloted by a crew consisting of V.M. Komarev, commander, K.P. Feoktistov, flight engineer, and B.B. Egorov, physician. The second flight, on Voskhod-2, was piloted by P.I. Belyayev, commander, and A.A. Leonov, copilot. During the second flight on Voskhod-2, on 18 March 1965, the first space walk (by A.A. Leonov) was completed.

In 1966 Center specialists began to train cosmonauts in a program involving testing and a series of flights on the new multiperson, multipurpose piloted spacecraft of the Soyuz series, which possessed technical capacities that significantly exceeded those of Vostok. One of the main goals of cosmonaut training here was to enable mastery of spacecraft approach and docking — key operations in the use of long-term orbital piloted stations designed for long periods of living and working in space.

At the same time (1966–1971) as cosmonauts were being prepared for piloted flights on Soyuz, the Center was providing training for flight programs on other types of spacecraft. Thus, between 1966 and 1968, cosmonauts were trained for long-term (up to 10 days) flights on Voskhod series spacecraft. In 1966, a separate group of cosmonauts was formed for training on the program for the Spiral aerospace system (device-50). Crew training on this program was discontinued in the early 1970s, and the cosmonauts in it were transferred to training for flights in the Soyuz and Soyuz-VI, as well as for the Salyut and Almaz space station programs.

During these years, the Center began actively training cosmonauts for translunar flights and lunar landing (lunar programs: 7K-L1, N1-L3) programs, which continued until the early 1970s. Work on the lunar programs is little known to the general public because these flights were never implemented. The group being trained for these programs included 15 military pilots, 7 specialists from industry, and 4 scientists from the USSR Academy of Sciences. A.A. Leonov and V.F. Bykovskiy were named the commanders of the first crews. Based on the

results of crew training, the Air Force command concluded that they were ready to complete a flight to the Moon.

During this same period, the Center was training cosmonauts on programs for orbital flights on Soyuz-7K-OK, Soyuz-VI-7K-VI, the Salyut long-duration space station, and the Almaz space station. Crew flights on the first long-duration space station, Salyut, began after it was inserted into orbit on 19 April 1971. After this, the Center's mission included training cosmonauts for long-duration spaceflights.

In 1972, it was decided to begin targeted training of crews for flights on the Almaz space station program, developed under the direction of the General Designer, Academician, V.N. Chelomey. In September 1972, ground-based tests of the Almaz station began, including tests of its thermal control and life support systems, which required long periods of confinement by the Center's cosmonauts and observers in a facility simulating the conditions of spaceflight. The first Almaz station was inserted into orbit in early 1973, but, after several days of unpiloted flight, it lost pressurization, and the crew that was all set to fly had its orders canceled. Crews began to be trained for flights on the next station.

The second Almaz manned station (launched as Salyut-3) was successfully inserted into orbit on 25 June 1974, and the third (called Salyut-5) began its flight on 22 June 1976. The Almaz program comprised five flights, and three crews lived and worked on the stations (one on Salyut-3 and two on Salyut-5). In 1977, the flights in this program were suspended. The official decision to cancel the Almaz program was not made for a number of years. Throughout this period, the Cosmonaut Training Center continued to train crews for the Almaz program until the early 1990s. Cosmonauts who had been assigned to this program were on "active standby" status until 1984. The flight of Salyut-5 marked the culmination of a major phase of work associated with the development, testing, and practical validation of the first generation of space stations.

While the Almaz stations were operating, the Center was also training crews for the Salyut program. Because the 3 April 1973 launch of the next station in this program, Salyut-2, was unsuccessful (all of the propellant the station carried was expended during the first orbital pass), crews designated for flight on Salyut were transferred to the international Apollo-Soyuz experimental program, based on a U.S.-U.S.S.R. agreement signed in May 1972) for a joint spaceflight. Four crews were formed for this flight. For the first time, instructors and cosmonauts at the Center had to train both their own and U.S. crews for a spaceflight at their own facility and at the U.S. Apollo-Soyuz base. During the training process, major difficulties arose as a result of the language barrier and the constraints imposed by the security restrictions of that time period.

On 26 December 1974 the Salyut-4 long-duration space station was launched; two crews lived and worked on it. The Apollo-Soyuz project culminated in the successful docking and joint flights of Soyuz-19 and Apollo 18 between 15 and 21 July 1975. The prototype of the first international manned space station was thus assembled in orbit.

Since that time, the Center has been directly involved in tasks related to cosmonaut training and piloted spaceflight implementation as entailed by international collaboration. This type of collaboration was continued in the Interkosmos program. In the fall of 1976, the first group of candidates for Interkosmos

program flights arrived for space station flight training in Star City from Czechoslovakia, East Germany, and Poland. A second group, which arrived in 1978, consisted of citizens of Bulgaria, Hungary, Vietnam, Cuba, Mongolia, and Rumania. These groups consisted primarily of pilots who had been trained in the Soviet Union's military academies or military flight schools. Interkosmos program flights began in 1978.

On 29 September 1977, the first of the new (second) generation space stations (Salyut-6) was inserted into orbit. The Center began plan-driven work to train cosmonauts to use this station for scientific research and industrial applications, especially those associated with the study of Earth in the interests of the economy and environmental protection. The fact that the station had two docking ports enabled a significant increase in the number of manned flights (because they made it possible for prime and visiting crews to work together on the station). This required intensifying the cosmonaut training process as well. The Center's success in solving problems related to cosmonaut training for long-duration spaceflights made it possible to set successive records for crew flight duration, of 96 days, 140 days, 175 days, and 185 days. Flights on Salyut-6 terminated in May 1981. A total of five prime crews had lived and worked on this station. These crews had been joined by 11 visiting crews. The Salyut-6 program included 16 piloted flights, of which 9 were Interkosmos flights. For its significant contribution to implementation of long-duration spaceflights and Interkosmos flights, the Center was awarded the Order of Friendship Among Peoples.

On 19 April 1982, the long-duration space station Salyut-7 was placed in orbit to replace Salyut-6. Six prime crews were trained for and flew on Salyut-7. Four visiting crews were trained for the station. Training for flights on Salyut-7 took place in 1986.

Starting in the second half of the 1970s, the Soviet Union actively worked on developing the Buran space shuttle. It was successfully tested in the unmanned mode on 15 November 1988 (launch from the Baykonur Cosmodrome, insertion into the predetermined orbit by the Energia booster rocket, a flight of two Earth orbits, prelanding maneuvers, and landing in automated mode on the landing strip at Baykonur). From 1979 up to the termination of this program in 1995, the Cosmonaut Training Center periodically trained (refresher course) cosmonaut candidates for Buran flights. A total of more than 20 people underwent such training.

The most intensive and large-scale training of cosmonauts was conducted as part of the 15-year implementation of the Mir Space Station program. The station's core module was inserted into orbit on 20 February 1986. The Mir station differed from previous stations by virtue of the fact that it had six docking ports, two along the long axis of the station and four perpendicular to it. The presence of six docking ports made it possible to assemble a scientific research complex in orbit from a number of modules, each of which had its own functional purpose. One of these modules had the capacity to dock with the U.S. Space Shuttle. From 8 September 1989 to 28 August 1999, the Mir station operated continuously in manned mode. The crews relieved each other in flight. The station was equipped with unique scientific instrumentation and experimental apparatus (more than 240 setups with a total weight of approximately 11.3 tons), which had been developed in 27 nations. Between 1986 and the moment the Mir station ceased to exist (23 March 2001), 28 prime crews were trained and flew on

that station. These crews had as members 35 Russian cosmonauts, 7 U.S. astronauts, and 1 crewmember each from the European Space Agency and France. In addition to the prime crews, Mir was visited by 16 visiting crews who remained from 1 week to 1 month; of these, 15 crews were part of a program of international collaboration with participants from Syria, Bulgaria, Afghanistan, France (5 crews), Japan, United Kingdom, Austria, Germany (2 crews), the European Space Agency, and Slovakia. A total of 104 cosmonauts and astronauts lived and worked on the station, including 62 foreign nationals (from 11 countries and the European Space Agency). These cosmonauts performed more than 23,000 scientific experiments and research studies on the station as part of national and international programs. In accordance with the Mir-NASA program, 9 visiting crews were transported to the station on Space Shuttles, including 37 U.S. astronauts. The total duration of the flights of foreign cosmonauts on Mir was 4.3 years, which represents 32% of the total time the station functioned in manned mode. The Mir Space Station served as a unique testing ground for developing of technological processes associated with training international crews, and for developing international cosmonaut training systems, which have subsequently been applied in the International Space Station program.

Starting in 1993, pursuant to Russia's international commitments to develop and use the International Space Station, the Center has been participating directly in training cosmonauts and astronauts to support the International Space Station Program.

Objectives, Structure, and Technical Facilities of the Yu.A. Gagarin Russian State Scientific Research and Testing Cosmonaut Training Center

As space technology developed, the conditions of spaceflight changed, more practical experience was gained, and the goals, structure, and technical facilities of the Center continually underwent adjustment and change. Now, the Center is organized into the following divisions: the cosmonaut corps, the professional and biomedical cosmonaut training, the scientific research and testing divisions and laboratories, the flight divisions, the prototype production divisions, a special separate aviation test and training regiment, and the support divisions.

These are the main areas of Center activity:

- to serve as the lead organization for selecting cosmonaut candidates and training cosmonauts in all categories and specialties for flights on any spacecraft, developing and improving cosmonaut training systems, medical support of cosmonaut training and cosmonaut postflight rehabilitation, and developing and deploying technical devices (technical facilities) for cosmonaut training;
- to provide general spaceflight training to cosmonaut candidates, to provide group and crew training, to support and upgrade trainee qualifications, and to train foreign astronauts (cosmonauts);
- to conduct fundamental, exploratory, and systems research, development work and scientific testing to improve the quality of cosmonaut training, the

efficiency of spacecraft crew performance, and spaceflight safety; and to advance spaceflight technology and the development of Center technical facilities;

- to participate in flight tests of spacecraft, ergonomic evaluations, and scientific and technical oversight tracking of the development and use of piloted spacecraft;
- to train and certify scientific work forces;
- to participate in developing of Russian Federation policies regarding space exploration, to draft proposals for inclusion in the Federal Space Program of Russia, to develop draft regulations on issues relating to space exploration, and to draft collaboration agreements with foreign countries in the area of space exploration, and to perform public relations functions publicizing Russia's achievements in spaceflight research and applications.

Cosmonaut candidates are selected from among Russian Federation citizens who have expressed the desire to participate in spaceflights and who meet the established professional and medical requirements. Individuals who pass the special selection process are accepted into the cosmonaut team of the Russian Federation (Center) and appointed to the position of candidate cosmonaut pilot or candidate mission specialist. Candidate cosmonaut pilots are selected from among fighter pilots, pilots, and other specialists who have received higher technical education in the area of space technology. Candidates for mission specialist are selected from among scientists, engineers, physicians, and military and other specialists who have academic training and appropriate professional skills for implementing scientific research and experimental programs on spaceflights.

When the professional selection system was developed, one of the initial tenets was that cosmonaut selection is a continuous process, based on a systematic evaluation of candidates' health, physical and psychological qualities and educational level by using a set of methods for social, educational, medical, and psychological selection. Social selection methods include analyzing and evaluating an individual's moral and ethical qualities, as well as motivations, interests, needs, intragroup relationships, resistance to social pressure, and ability to adapt to a new social environment. Educational selection involves analyzing and evaluating the level of knowledge, skills, and professional experience a candidate has acquired to prepare for learning the chosen profession of cosmonaut. Medical selection is directed at identifying individuals whose health status and physical fitness would permit them to function as cosmonauts. Psychological selection methods are intended to identify the presence and strength of the set of personality traits that are required for becoming a cosmonaut and are conducive to successful mastery of this profession.

Cosmonaut training for flights involves three phases:

- general spaceflight training for cosmonaut candidates;
- training cosmonauts in groups based on specialization, types of piloted spacecraft, or area of activity;
- training designated cosmonaut crews (prime, standby, and backup) for spaceflight on specific piloted spacecraft.

Cosmonauts on the cosmonaut team, who have not been included in a group or crew and whose health status meets the standard requirements, periodically undergo short refresher courses according to a specially developed training curriculum to maintain their knowledge, skills and abilities at the requisite level. The major objective of general spaceflight training for cosmonaut candidates is to ensure their mastery of the principles of cosmonautics and related fields and their acquisition of the capacity to become cosmonaut pilots or mission specialists. The major objective of cosmonaut group training is to improve their professional qualities, their specialization in particular types of spacecraft or areas of work, monitor their status and maintain their health, and also maintain high performance capacity. The major objective of cosmonaut crew training is to ensure that crews are prepared to complete the flight program on a specific spacecraft. This sequence of training phases is mandatory for everyone. Candidates receive general spaceflight training once; however, they may pass through the other training phases more than once.

These are the goals of the first cosmonaut training phase:

- to learn the theoretical principles underlying cosmonautics;
- to study the design and principles of construction of the basic spacecraft, its utility, and special-purpose systems and equipment;
- to gain an overview of systems on a piloted spacecraft;
- to learn the principles for conducting tests, research, and experiments on board a piloted spacecraft;
- to learn about international legal standards for the use of space;
- to gain experience with the effects of dynamic spaceflight factors;
- to acquire theoretical knowledge and the elementary practical skills for working in spacesuits;
- to acquire theoretical knowledge and elementary practical skills for EVAs;
- to develop practical skills for landing in various extreme climatic or geographical areas;
- to undergo flight and parachute training;
- to undergo a series of biomedical measures and exercises directed at monitoring status, maintaining and strengthening health, and identifying the individual psychophysiological characteristics of each cosmonaut candidate;
- to learn how to ensure the safety of spaceflight.

These are the goals of the second phase of cosmonaut training:

- to study the design, layout, and onboard systems of specific piloted spacecraft;
- to develop skills needed to work with onboard systems and scientific equipment;
- to develop skills needed to perform flight procedures and the components of the flight program;
- to master generic operations for EVA;

- to develop skills for performing generic operations involved in technical maintenance, repair, and inventorying of equipment;
- to undergo medical measures directed at maintaining cosmonauts' good physical condition and functional psychophysiological capacities, high performance level for professional tasks, and practical skills for diagnosing illnesses and conditions and providing medical aid to themselves and other crew members.

These are the goals of the third phase of cosmonaut training:

- to study the design characteristics of a specific piloted spacecraft, its systems, scientific and specialized equipment, and also guidelines for using them;
- to study the program for the scheduled flight, onboard documentation and documents regulating crew interactions with ground control and flight support groups.
- to improve skills and develop abilities to control a spacecraft and use its systems and equipment during all flight phases in standard and contingency modes;
- to gain practical mastery of flight program components;
- to acquire experience with interactions among crew members and between the crew and ground control groups;
- to undergo medical measures directed at ensuring a state of good health, high crew performance capacity, and readiness to perform the biomedical section and flight program as a whole.

Training of cosmonaut candidates and cosmonauts has the following forms:

- study of theoretical disciplines, designs, and the principles of construction of a piloted spacecraft and its utility and special purpose systems and equipment;
- practical exercises and training on training simulators of various types and purposes, aircraft, test stands, mock-ups, setups, airborne laboratories, underwater laboratories, centrifuges, and barochambers;
- participation in tests of space technology, research, and performance of expert evaluations (Figs. 2 and 3).

The list of types of cosmonaut training (relevant to all phases) includes

- technical training;
- comprehensive and specialized training on simulators and test stands for teaching control of spacecraft and transport vehicles and use of onboard systems and equipment;
- flight and special parachute training;
- biomedical training;
- training under simulated spaceflight conditions;
- mastery of the set of life support system components for cosmonauts in an artificial environment;



Figure 2. Soyuz Transport Spacecraft mock-up. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

- comprehensive training in the EVA program;
- training in performing applied scientific research and experiments;
- training in navigational and ballistic support of spaceflight (space navigation);
- training to use movie cameras, television cameras, and to work with the onboard videotape system;
- training in what to do upon landing in various extreme climatic and geographic conditions;
- training in the operations involved in technical maintenance and repair of spacecraft and their onboard systems;
- training in controlling space robotic systems and manipulators;
- training in onboard and flight documentation, flight programs, and rules for structuring crew work on board;
- training to perform experimental work and testing of spacecraft and their systems;
- training in what to do in contingency (emergency) situations;

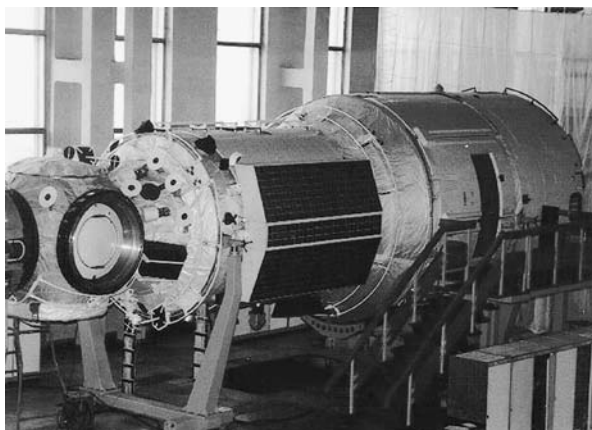


Figure 3. Mock-up of Russian section of the International Space Station. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

- psychological training;
- humanitarian training (legal, linguistic, culture-specific training, and training for symbolic activity).

All phases of cosmonaut candidate and cosmonaut training are held at the Center facilities. If necessary for training cosmonauts, the technical devices of other organizations and businesses are brought in on a contractual basis.

The Center's facilities consist of equipment, which can be grouped on the basis of their function or purpose. These are the most important:

1. Comprehensive trainers simulating transport spacecraft and space station modules to foster mastery of all flight programs on transport spacecraft and space stations, under standard conditions and in various contingency and emergency situations, including performance of repairs.
2. Specialized trainers simulating transport spacecraft and space station modules to simulate individual phases, operations, and modes of flight (trainers simulating approach, berthing and docking the spacecraft with an orbital station, including remote control and manual control of the trajectory for atmospheric reentry using a centrifuge, control of robotic, space-suit, life support, and airlock decompression systems).
3. Simulators of space and spaceflight conditions.

- CF-18 centrifuge (18-meter rotational arm, maximum G-force of 30 G, maximum G-force gradient of 5 G/s, replaceable cabins simulating spacecraft work stations), designed for physiological studies and cosmonaut (pilot) training in tolerance of G-forces controlled with respect to magnitude and direction under conditions of altered microclimate (pressure, temperature, humidity, and atmospheric composition), and for testing aerospace technology and performing experiments (a specialized dynamic trainer simulating

manual control of reentry of the Soyuz TMA spacecraft with deceleration conditions created by the CF-18 centrifuge) (Fig. 4);

- CF-7 centrifuge (7-meter rotational arm, maximum G-force of 10 G, maximum G-force gradient of 7 G/s, replaceable work station in the cabin), designed for training and study of human tolerance of various types of G-force (a specialized dynamic trainer simulating manual control of reentry of the Soyuz TMA spacecraft under exposure to deceleration created by the CF-7 centrifuge base);
- underwater laboratory (pool diameter of 23 meters, height of 12 meters, number of windows: 45, capacity of platform on which mock-ups are placed: up to 8 tons), for training cosmonauts for EVAs under conditions of simulated weightlessness in an underwater environment and for developing methods for training cosmonauts and testing space technology (Fig. 5);
- aircraft laboratories:
 - (a) Il-76MDK-2, designed for flights creating short periods of weightlessness and decreased weight for training spacecraft crews, conducting biomedical research, and testing equipment (can create weightlessness lasting up to 28 seconds; on one flight can generate periods of weightlessness up to 20 times during a flight of 1.5–2 hours) (Fig. 6);
 - (b) Tu-154M-LK-1 and Tu-134LK, designed for training cosmonauts on navigation and orientation and conducting research on natural and artificial objects on Earth's surface using remote sensing) (Fig. 7);
- outer space planetarium for classes and training on the stars in the sky and development of cosmonaut skills in celestial navigation and celestial orientation (the planetarium reproduces 9000 stars and makes it possible to simulate stars at altitudes up to 500 kilometers with a variation of angular velocities of rotation of the star field in increments of 0.002–5.0°/s with a reproduction error of angular distances between stars of 12 arc minutes);
- “Dry Land” and “Ocean” special training simulators (specially equipped mock-ups of reentry modules) for training cosmonauts to survive after emergency reentry module landings in various climatic and geographic areas;

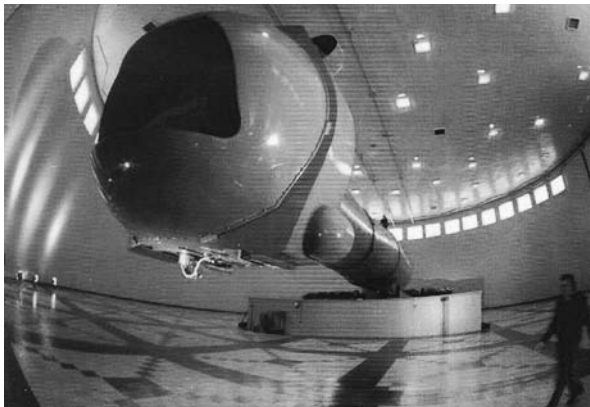


Figure 4. CF-18 centrifuge. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

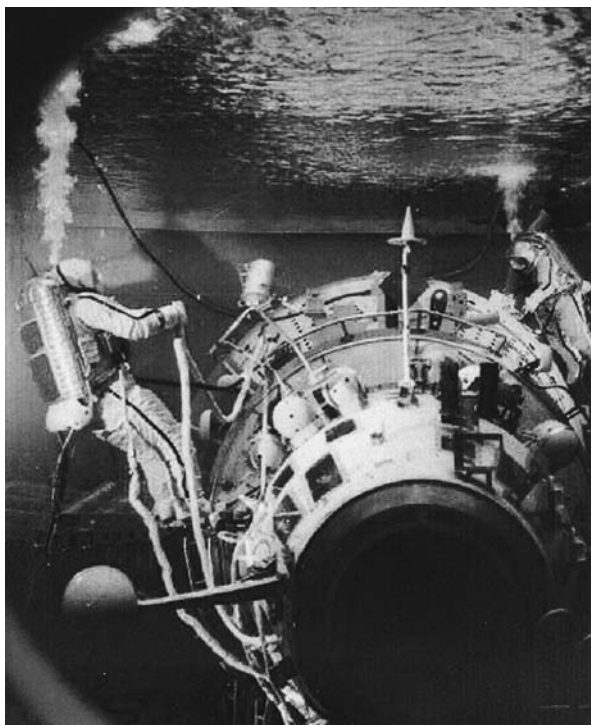


Figure 5. Cosmonaut training in an underwater environment. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

- a set of instruments and equipment for biomedical cosmonaut training;
- means for flight (training aircraft) and special parachute training;
- SBK-80 stationary barochamber (the main chamber has a volume of 7.5 m^3 with work space for four people; the pressure differential chamber is 1.8 m^3 in volume with work space for one; the maximum atmospheric rarefaction created is 41 mmHg, which corresponds to 20 km above sea level, the ascent time to 12 km is 100 seconds, and the emergency altitude decrease time from 20 km to 0 kilometers is not less than 10 seconds), designed for altitude training of cosmonauts under ground-based conditions to evaluate their physiological tolerance of rarefied atmospheres, and for conducting medical research and testing of life support systems.

The training simulators for the Russian portion of the International Space Station are a flexible, expandable set of programmable hardware devices for various training configurations developed on the basis of crew training tasks, as well as the evolution of the International Space Station in the process of its assembly and use. The training complex for the modules of the Russian station segment are being created phase by phase (add-on method) as the flight prototypes are developed.

The devices used for cosmonaut training also include actual space technology and apparatus prototypes, life size mock-ups of spacecraft, stations and their



Figure 6. Cosmonaut training under conditions of simulated weightlessness in an aircraft. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

modules, various research and functional simulation stands, a training simulator facility and devices for physical and medical cosmonaut training, and reproduction of equipment and communications devices using personal computer training systems.

In addition to Center equipment and facilities, cosmonaut training may use equipment belonging to other organizations and departments (the test stand facilities of organizations that develop piloted spacecraft and their systems, equipment belonging to the Cosmodrome and the Flight Control Center, search and rescue equipment, Naval vessels and Air Force helicopters, and equipment belonging to experimental and research organizations). International collaboration has required that training also use equipment from foreign countries and international organizations participating in joint international space programs.

Center activity is not limited to training cosmonauts for flight. During the 30 years of its existence, the Center has performed more than 1500 tests of various devices. The Center has participated in 93 flight tests and development tests of piloted spacecraft and conducted 260 tests of cosmonaut training technology and various other equipment. Approximately 450 testing and research



Figure 7. Cosmonaut training after reentry module landing. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

studies were performed on flying laboratories under conditions of simulated weightlessness (involving more than 1600 flights, during which weightlessness was produced approximately 14,000 times).

ALEXANDER N. EGOROV

YU.B. SOSYURKA

V.I. YAROLOV

Yu.A. Gagarin Cosmonaut Training Center
Russia

D

DEEP SPACE NETWORK, EVOLUTION OF TECHNOLOGY

The Deep Space Network

The Deep Space Network (DSN), managed by the California Institute of Technology's Jet Propulsion Laboratory (JPL), has provided vital communications and navigation services for NASA deep space exploration missions for more than 40 years. The remarkable technical achievements of the planetary exploration program executed by NASA would not have been possible without the extensive and sophisticated communications and data handling systems that comprise the Deep Space Network. The high-resolution pictures of Jupiter's satellites that were transmitted by the Galileo spacecraft are an example of this. Another is the ability to continue communicating with a spacecraft that was launched 30 years ago and is now more than 7 billion miles from Earth (Pioneer 10). These capabilities are the result of a sustained effort to improve the data rate capability and the sensitivity of the receivers by the research and development staff of the Jet Propulsion Laboratory in Pasadena California.

The principal facilities of the DSN are three major ground station complexes, one in the United States (Goldstone, California), one near Madrid, Spain, and one near Canberra, Australia. Each of the complexes has several tracking antennae; the largest has a diameter of 70 meters, and smaller ones have diameters of 11–34 meters. In addition to antennae, the DSN also has a complex of computers and signal processing capabilities that permit very thorough and sophisticated analysis of signals sent back from distant spacecraft.

In addition to receiving data from very distant spacecraft, the facilities of the network are also used to control the spacecraft. The network and the associated spacecraft are designed so that if a failure of a spacecraft should occur,

there are often means available to effect a recovery. Perhaps the best example of this is the work-around that was developed as a result of the failure of the high-gain antenna on the Galileo spacecraft. In spite of this setback, the Galileo mission was judged more than 70% successful because of the measures taken to adapt the data receivers and the processors to the failure. Finally, the continuing improvements in the technology of the network have greatly increased the useful lives and capabilities of the spacecraft that use the network. The example of Pioneer 10 has already been mentioned. It was originally designed for a life of about 5 years, and because of continued technical improvements in the network, can still be tracked by the network 30 years later.

A thorough history of the DSN is available in a recent published book *Uplink-Downlink: A History of the Deep Space Network* (1) in which these achievements are described in detail.

No history of the DSN would be complete without full appreciation of the contribution made by advanced technology to the successful development of the Network. The wellspring of new and innovative ideas for increasing the existing capability of the Network, improving reliability, operability, and cost-effectiveness, and for enabling recovery from potential mission-threatening situations has resided, from the very beginning of the Network's history, in a strong program of advanced technology, research, and development. Many of the accomplishments of the technology program, known during much of this time as the DSN Advanced Systems Program, can be found in Ref. 2.

The Great Antennae of the Deep Space Network

To enable 24-hour coverage of deep space spacecraft, NASA/JPL maintains a complex of large antennae at each of three geographic locations; Goldstone, California, Madrid, Spain, and Canberra, Australia. A photograph of the NASA Deep Space Communications Complex (DSCC) near Canberra, Australia, is shown in Fig. 1.

The large antennae at the complexes are quasi-parabolic reflector antennae; one has a diameter of 70 meters, the others are 34 meters in diameter. These are used for deep space mission support. Other smaller (26-m or 11-m) antennae provide occasional tracking support for selected Earth-orbiting missions.

Each of the antennae has what is termed a Cassegrain configuration and has a secondary reflector mounted on the center axis just below the focal point of the primary reflector "dish." The secondary reflector relocates the antenna focal point closer to the surface of the main dish and thus establishes a more convenient location for low-noise amplifiers, receivers, and powerful transmitters.

The efficiency with which parabolic antennas collect radio signals from distant spacecraft is degraded to some extent by radio noise radiated by Earth's terrain surrounding the antenna. This form of radio noise, known scientifically as "blackbody radiation," is a physical characteristic of all material that have a temperature above absolute zero (-273°F . or 0K). The magnitude of noise power radiated by a material body depends on its temperature. It is incredibly small at the temperature of typical Earth surfaces, but it is enormous at the temperature of the Sun, for instance. All parabolic, radio antennae have side



Figure 1. Deep Space Communications Complex; Canberra, Australia. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/ant_1.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

lobes in their beam patterns, and the magnitudes of those side lobes can increase as the antennae deviate from the ideal shape. The shape of the beam and its “side lobes” are essentially the same whether the antenna is used for transmitting or receiving signals. The “side lobes” are analogous to the circles of light surrounding the main beam of a flashlight when it is held close to a reflecting surface.

When an antenna is used for transmitting a strong signal to a spacecraft, the side lobes are of no great consequence. However, when the antenna is receiving a weak signal from a distant spacecraft, particularly at or near the horizon, “Earth noise” picked up by the side lobes can be sufficient to obscure the spacecraft signal in the extremely sensitive receivers used on the DSN antennae. This produces errors in the data stream being delivered to spacecraft engineers and scientists and may cause the DSN receivers and antennae to lose the spacecraft signal altogether; in that case, the data stream is completely lost.

The earliest antennae in the DSN were of commercial design and were parabolic. Then, as now, the actual efficiency of the antenna represented a compromise between maximum signal gathering capability and minimum susceptibility to radio noise picked up from surrounding Earth.

Improved technology to reduce the side lobes and increase the signal collection capability (gain) of future DSN antennae appeared in the DSN’s Advanced Systems Program in the early 1970s. The new technology was based on a “dual-shape” design wherein the surface shapes of both the primary and secondary reflectors were modified to illuminate the slightly reshaped “quasi-parabolic” surface of the main reflector more uniformly (3). However, it was not until the 1980s when the first 34-meter high-efficiency antennae were built that the new “dual-shape” design saw operational service in the Network.

These antennae were needed by the DSN to support the two Voyager spacecraft in their tour of the outer planets. At the time, the DSN was in transition from the lower, less capable, S-band (2.3 GHz) operating frequency for

which the early spacecraft and antennae were designed, to the X band (8.4 GHz), a higher, more capable operating frequency. When X-band technology became available (1975) in the DSN, all later spacecraft and DSN antennae were designed to operate at X-band frequencies. These antennae were, therefore, the first to be optimized for best performance in the X band.

As the Voyager 2 spacecraft headed outward toward Neptune, it was recognized that increased signal collecting area was needed on Earth to support this unique scientific opportunity effectively. The DSN's largest antennae at the time were 64-m parabolas of the original design. Calculations showed that the best investment of scarce construction funds would be to modify these antennas using the dual-shape design and expanding their diameter to 70 m. It was also apparent that the upgraded large antennas would benefit the planned Galileo and Magellan missions.

Completed in time for support of Voyager 2 at Neptune, the 70-m enhancement project (Fig. 2) resulted in an increase of more than 60% in the effective collecting area of these large antennae. Fully half of the increase was attributed to the dual-shape design.

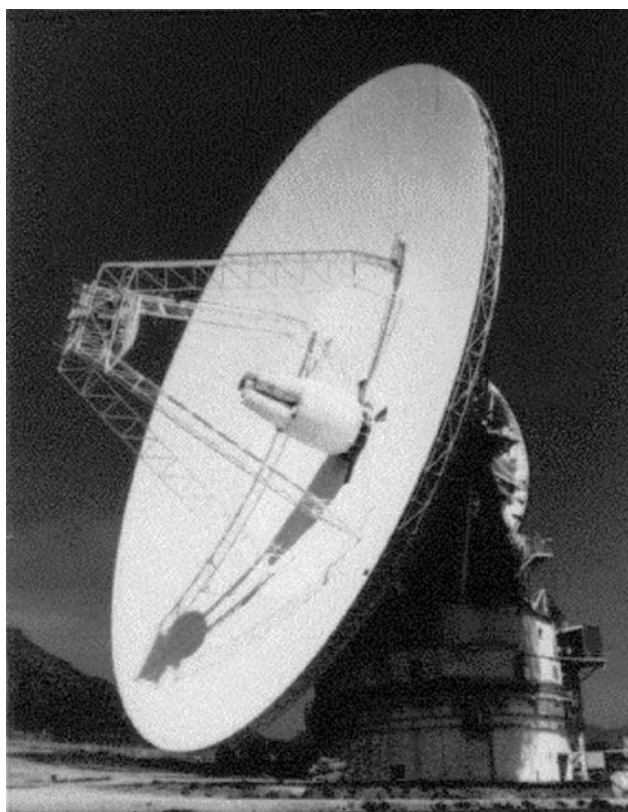


Figure 2. The 70-meter antenna with dual-shaped reflector design. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/ant_4.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

By the end of the century, several new 34-m antennae employing the dual-shaped reflector design in conjunction with beam waveguide (BWG) techniques had been constructed for operational use in the Network. The dual-shaped reflector design enhanced the radio performance of the antenna, whereas the beam waveguide configuration greatly facilitated maintaining and operating of the microwave receivers and transmitters. Using a series of additional secondary reflectors to relocate the focal point into a stationary room below the main dish, the BWG design feature enabled mounting these critical components in a fixed environment rather than in the more conventional type of moving and tipping enclosure mounted on the antenna itself.

Beam waveguide antennae had been used for many years in Earth communications satellite terminals where ease of maintenance and operation outweighed the consideration of losses introduced into the microwave signal path by the additional microwave reflecting mirrors. For deep space applications, however, where received signal power levels were orders of magnitude smaller, any losses in the signal path were a matter of great concern, and the losses associated with BWG designs kept such antennas out of consideration for DSN purposes for many years. Researchers in the DSN Advanced Systems Program, nevertheless, pursued the idea of BWG antennas for the DSN and by 1985 were ready to conduct a collaborative experiment with the Japanese Institute for Space and Aeronautical Sciences (ISAS), using its new 64-m beam waveguide antenna at Usuda, Japan (4). Using one of the DSN's low-noise microwave receivers installed on the Usuda antenna to receive a signal from the International Cometary Explorer (ICE) spacecraft, the researchers made very precise measurements of the microwave losses, or degradation, of the downlink signal.

The results of the experiment were very surprising. The measured losses attributable to the BWG design were much smaller than expected, exhibited similar performance at zenith and better performance at low elevation angles than traditional antennae, and confirmed the efficacy of the BWG configuration.

Encouraged by this field demonstration, researchers sponsored by the Advanced Systems Program moved forward with the construction of a prototype BWG antenna for potential application in the Network.

This new prototype BWG antenna, built at the Venus site at Goldstone, replaced an aging 26-m antenna that had served for many years as a field test site for technology research and development (R&D) programs. The designers used microwave optics analysis software, an evolving product of the technology program, to optimize the antenna for operation across a wide range of current and future DSN operating frequency bands. When completed, the antenna successfully demonstrated its ability to operate effectively in the S band, X band, and Ka band (approximately 2, 8, and 32 GHz, respectively).

Figure 3 shows the completed BWG antenna, and Fig. 4 shows the interior of the equipment room below the antenna structure.

The various frequencies and modes of operation for the BWG antenna were selected by rotating the single microwave mirror at the center of this room. Lessons learned by Advanced Systems Program personnel in constructing and evaluating this antenna were incorporated into the design of the operational BWG antennae for the rest of the Network; the result was that their performance somewhat exceeded that of the prototype, especially at lower frequencies.



Figure 3. The 34-meter beam waveguide antenna at DSS 13. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/ant_5.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Forward Command/Data Link (Uplink)

The large antennae of the DSN are used for transmitting radio signals carrying instructions and data to the spacecraft, as well as for receiving signals back from them. Getting data to distant spacecraft safely and successfully requires



Figure 4. Stationary equipment room below BWG Antenna. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/ant_6.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

transmitting substantial power from the ground, directed in a narrow beam toward the spacecraft. For most “normal” situations, the compatible design of spacecraft and the DSN is such that power of about 2 to 20 kW is adequate. However, situations in space are not always normal. Unexpected events can redirect a spacecraft’s main antenna away from Earth, leaving only a low-gain or omnidirectional antenna capable of receiving anything from Earth. Transmitter power up to 400 kW in the S band can be sent from the 70-m antenna during attempts to regain contact with a spacecraft in an emergency.

The initial design and evaluation of R&D models of the high-power transmitters and their associated instrumentation were carried out under the Advanced Systems Program. Much of the essential field testing was carried out as part of the planetary radar experiments. This cooperative and productive arrangement provided a realistic environment for testing without exposing an in-flight spacecraft to an operationally unqualified uplink transmission. Later, DSN engineers implemented fully qualified operational versions of these transmitters in the Network at all sites.

Pointing the narrow forward link signal to the spacecraft is critical, especially when making initial contact without having received a signal for reference, as is typical of emergencies. The beam width of the signal from the 70-m antenna in the S band is about 0.030° , and that of the 34-m antenna in the X band is about 0.017° . Achieving blind pointing to that precision requires thorough understanding of the mechanics of the antenna, including the effects of gravity and wind on the dish, specifics of the antenna bearing and positioning mechanisms, as well as knowledge of the spacecraft and antenna positions, atmospheric refraction, and other interferences.

Forward link data delivered to a spacecraft, if incorrectly interpreted, may cause that spacecraft to take undesirable actions, including some that could result in an emergency for the spacecraft. To guard against that possibility, the forward link signal is coded with additional redundant data that allow the spacecraft data system to detect or correct any corruption in that signal. Operating on the presumption that it is always better to take no action than an erroneous one, the forward link decoder accepts only data sets for which the probability of error is extremely small and discards those that cannot be trusted (5).

Return Telemetry/Data Link (Downlink)

Throughout the Network, the stations use the same antennae for both the forward link and the return data link signals. Because the strength of a signal decreases as the square of the distance it must travel, these two signals may differ in strength by a factor of 10^{24} in a single DSN antenna. Isolating the return signal path from interference by the much stronger forward signal poses a significant technical challenge. Normally, these two signals differ somewhat in frequency, so at least a part of this isolation can be accomplished via dichroic or frequency selective reflectors. These reflectors consist of periodic arrays of metallic/dielectric elements tuned for the specific frequencies that either reflect or pass the incident radiation. These devices must be frequency selective, and they also must be

designed to minimize the addition of extraneous radio noise picked up from the antenna and its surroundings, which would corrupt the incredibly weak signals collected by the antenna from the desired radio source in deep space.

The DSN Advanced Systems Program developed the prototypes for almost all the reflectors of this type currently used in the Network. As an adjunct to this work, powerful microwave analytical tools that can be used to affix design details for almost any conceivable dichroic reflector applicable to the frequency bands of the DSN were also developed under the Program.

Low-Noise Amplifiers. The typical return data link signal is incredibly small and must be amplified before it can be processed and the data itself reconstructed. The low-noise amplifiers that reside in the antennae of the DSN are the most sophisticated in the world and provide this amplification while adding the least amount of noise of any other such devices.

Known as traveling-wave masers (TWMs), the quietest (in adding radio noise) of these operational devices amplify signals that are propagated along the length of a tuned, ruby crystal. Noise in a TWM depends on the physical temperature of the crystal; those in the DSN operate in a liquid helium bath at 4.2 K. (Zero K is equivalent to a temperature of -273.18°C . Therefore, the temperature of the helium bath is equivalent to approximately -269°C) Invented by researchers at the University of Michigan, early development of practical amplifiers for the DSN was carried out under the DSN Advanced Systems Program (6). The quietest amplifiers in the world today (Fig. 5), which operate at a temperature of 1.2 K, were developed by the DSN Advanced Systems Program and demonstrated at the Technology Development Field Test Site, DSS 13 (7).

Some of the low-noise amplifiers in the DSN today are not TWMs but are a special kind of transistor amplifier using high-electron mobility transistors (HEMTs) in amplifiers cooled to a temperature of about 15 K (8).

Developed initially at the University of California at Berkeley, these amplifiers were quickly adopted by the scientific community for radio astronomy applications. This, in turn, spawned the JPL development work that was carried out via collaboration involving JPL, radio astronomers at the National Radio Astronomy Observatory (NRAO), and device developers at General Electric (9,10). This work built on progress in the commercial sector with uncooled transistor amplifiers. In the 2-GHz DSN band, the cooled HEMT amplifiers were almost as noise-free as the corresponding TWMs, and the refrigeration equipment needed to cool the HEMTs to 15 K was much less troublesome than that for TWMs. Primarily for this reason, current development efforts in the DSN are focused on improving the noise performance of HEMT amplifiers for higher DSN frequency bands.

The first DSN application of cooled HEMT amplifiers came with outfitting the NRAO Very Large Array (VLA) in Socorro, New Mexico, for collaborative support of the Voyager-Neptune encounter (9). The VLA, designed to map radio emissions from distant stars and galaxies, consists of 27 antennas, each 25 m in diameter, arranged in a triaxial configuration. Within the funding constraints, only a small part of the VLA could be outfitted with TWMs, whereas HEMTs for the entire array were affordable and were expected to give an equivalent sensitivity for the combined full array. In actuality, technical progress with the HEMTs during the several years taken to build and deploy the needed X-band

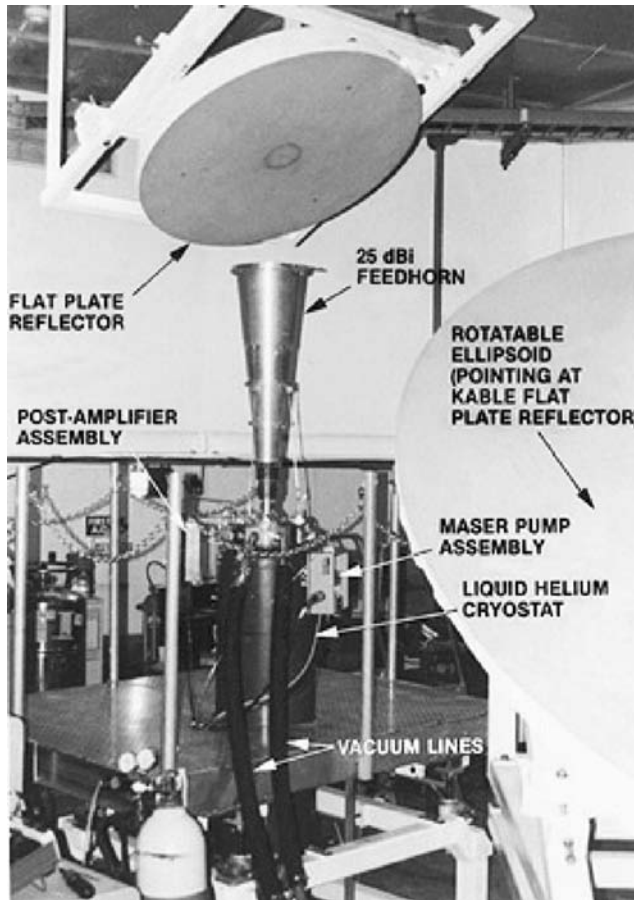


Figure 5. Ultra-low-noise amplifier at DSS 13. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/return_2.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

(8 GHz) amplifiers resulted in better performance of the fully equipped VLA than would have been possible with the VLA partially equipped with more expensive TWMs. Since that time, many of the DSN operational antennae have had cooled HEMT amplifiers installed for the 2- and 8-GHz bands.

Phase-Lock Tracking. Once through the first stages of processing in low-noise amplifiers, there are still many transformations needed to convert a radio signal from a spacecraft into a replica of the data stream originating on that spacecraft. Some of these transformations are by nature analog and linear, and others are digital with discrete quantization. All must be performed with virtually no loss in fidelity in the resultant data stream.

Typically, the downlink signal consists of a narrowband “residual carrier” sine wave, together with a symmetrical pair of modulation sidebands, each of which carries a replica of the spacecraft data. (Specifics of the signal values vary greatly but are not essential for this general discussion.) If this signal is

cross-correlated with a pure identical copy of the residual carrier, the two sidebands will fold together, creating a baseband signal that contains a cleaner replica of the spacecraft data than either sideband alone. Of course, such a pure copy of the carrier signal does not exist, but must be created, typically via an adaptive narrowband filter known as a phase-locked loop (11). The re-created carrier reference is thus used to extract the sidebands. The strength of the resultant data signal is diminished to the extent that this local carrier reference fails to be an identical copy of the received residual carrier. Noise in the spectral neighborhood of the received residual carrier and dynamic variations in the phase of the carrier itself limit the ability to phase lock the local reference to it.

These dynamic variations are dominated by the Doppler effect due to the relative motion between a distant spacecraft and the DSN antenna on the surface of spinning Earth. Over the years, the DSN Advanced Systems Program has contributed significantly to the theory and design practices for phase-locked loops and to the impacts of imperfect reference tracking on phase-coherent communications (12,13).

Synchronization and Detection. Further steps in converting a spacecraft signal into a replica of the spacecraft data stream are accomplished by averaging the signal across brief intervals of time that correspond to each symbol (or bit) transmitted from the spacecraft and by sampling these averages to create a sequence of numbers, often referred to as a “symbol stream.” These averages must be precisely synchronized with the transitions in the signal as sent from the spacecraft, so that each contains as much as possible of the desired symbol and as little as possible of the adjacent ones. Usually, a subcarrier, or secondary carrier, is employed to shape the spectrum of the spacecraft signal, and it must be phase-locked and removed before final processing of the data themselves. The Network contains several different generations of equipment that perform this stage of processing. Designs for all of these have their roots in the products of the DSN Advanced Systems Program. The oldest equipment is of a design developed in the late 1960s. This equipment was mostly analog and, while still effective, was subject to component value shifts with time and temperature and thus, required periodic tending and adjustments to maintain desired performance.

As digital devices became faster and more complex, it became possible to develop digital equipment that could perform this stage of signal processing. Digital demodulation techniques were demonstrated by the Advanced Systems Program in the early 1970s in an all-digital ranging system. Similar techniques were subsequently employed for data detection in the second generation of the Demodulator-Synchronizer Assembly (14,15).

A Digital Receiver. Rapidly evolving digital technology in the 1980s led researchers to explore the application of digital techniques to various complex processes found in receiving systems such as those used in the Network. The processes of filtering, detection, and phase-lock carrier tracking, formerly based on analog techniques, were prime candidates for the new digital technology.

In this context, the Advanced Systems Program supported the development of an all-digital receiver for Network use. Known as the Advanced Receiver (ARX), the developmental model embodied most of these new ideas and demonstrated capabilities far exceeding those of the conventional analog receivers then installed throughout the Network (16,17).

Encouraged by the performance of the laboratory model, an engineering prototype was built and installed for evaluation in an operational environment at the Canberra, Australia, tracking complex. Tests with the very weak signal from Pioneer 10 spacecraft, then approaching the limits of the current DSN receiving capability, confirmed the design's performance and significantly extended the working life of that spacecraft.

As a result of these tests, the DSN decided to implement a new operational receiver for the Network that would be based on the design techniques demonstrated by the ARX. The new operational equipment, designated the Block V receiver (BVR), would include all of the functions of the existing receiver in addition to several other data processing functions, such as demodulation and synchronization, formerly carried out in separate units.

As the older generation receivers were replaced with all-digital BVR equipment, the Network generally improved weak signal tracking performance and operational reliability. The receiver replacement program was completed throughout the Network by 1998.

Encoding and Decoding. Data generated by scientific instruments must be reliably communicated from the spacecraft to the ground, despite the fact that the signal received is extremely weak and the ground receiver corrupts the signal with additive noise. Even with optimum integration and threshold detection, individual bits usually do not have adequate signal energy to ensure error-free decisions. To overcome this problem, structured redundancy (channel encoding) is added to the data-bit stream at the spacecraft. Despite the fact that individual "symbols" resulting from this encoding have even less energy at the receiver, the overall contextual information, used properly in the decoding process on the ground, results in more reliable detection of the original data stream.

High-performance codes to be used for reliable data transfer from spacecraft to DSN were identified by research performed under the Advanced Systems Program and adopted for standard use in the Network, while the search for even more powerful and efficient codes continued. New, more efficient, block codes that used the limited spacecraft transmitter power better by avoiding the need to transmit separate synchronizing signals were developed and first demonstrated on the Mariner 6 and 7 spacecraft in 1969 (18). By putting the extra available spacecraft transmitter power into the data-carrying signal, the new block code enabled the return of Mars imaging data at the astonishing (for the time) rate of 16,200 bits per second, an enormous improvement over the 270 bits per second data rate for which the basic mission had been designed. Of course, conversion of the encoded data stream back to its original error-free form required a special decoder. The experimental block decoder developed for this purpose formed the basis for the operational block-decoders implemented in the Network as part of the Multimission Telemetry System, shortly thereafter (19).

While the JPL designers of the Mariner spacecraft were pursuing the advantages of block-coded data, the designers of the Pioneer 9 spacecraft at the NASA Ames Research Center (ARC) were looking to very complex convolutional codes to satisfy their scientists. The scientists agreed to accept intermittent gaps in the data, caused by decoding failure in exchange for the knowledge that successfully decoded data would be virtually error-free. In theory, a convolutional

code of length $k = 25$ would meet the requirement, but it had a most significant drawback. The decoding process was (at the time) extraordinarily difficult.

Known technically as “sequential decoding,” this was a continuous decoding operation rather than the “one block at a time” process used by the DSN for decoding Voyager data.

Originally, it was planned to perform the decoding operation for Pioneer 9 in nonreal time at ARC, using tape-recorded data provided by the DSN. However, Pioneer engineers, working in conjunction with the DSN Advanced Systems Program, explored and demonstrated the potential for decoding this code in real time via a very high-speed engineering model sequential decoder (20). The rapid evolution in the capability of small computers made it apparent that decoding Pioneer’s data in such computers was both feasible and economical. Subsequent implementation of sequential decoding in the Network was done via microprogramming of a small computer, guided by the knowledge gained via the efforts of the technology development program. The subsequent Pioneer 10 and Pioneer 11 spacecraft flew with a related code of length $k = 32$ and were supported by the DSN in a computer-based decoder.

The DSN standard code, flown on Voyager and Galileo, consisted of a short convolutional code that was combined with a large block-size Reed–Solomon code (21,22). The standard algorithm for decoding convolutional codes was devised in consultation with JPL researchers and demonstrated by simulations. Prototypes of the decoding equipment were fabricated and subsequently demonstrated at JPL.

The application of coding and decoding technology in the DSN was paced by the evolution of digital processing capability. At the time of the Voyager design, a convolutional code of length $k = 7$ was chosen as a compromise between performance and decoding complexity that would grow exponentially with code length. Equipment was implemented around the DSN to handle this code for Voyager and subsequently for Magellan, Galileo, and others. But modern digital technology permitted constructing much more complex decoders, so a code of length $k = 15$ was devised (23). This code was installed as an experiment on the Galileo spacecraft shortly before its launch. The corresponding prototype decoder was completed soon afterward. Though not used for Galileo because of its antenna problem, the more complex decoder was implemented around the Network to support the Cassini and subsequent missions.

Research in the technology development program provided the theoretical understanding to predict the performance of these new codes. Figure 6 displays the reliability of the communication error (actually, the probability of erroneous data bits), as it depends upon the spacecraft signal energy allocated to each data bit for uncoded communication and several different codes. In Fig. 6, the first set of curves shows the Voyager $k = 7$ code, both alone and in combination with the Reed–Solomon code. The second set of codes illustrates the $k = 15$ code, which was to be demonstrated with the Galileo’s original high-rate channel, shown alone and in combination with the Reed–Solomon code, either as constrained by the Galileo spacecraft data system ($I = 2$), or in ideal combination. The third set shows the $k = 14$ code, devised by the Advanced Systems Program researchers for the actual Galileo low-rate mission, both alone and in combination with the selected variable-redundancy Reed–Solomon code and a complex four-stage decoder. The

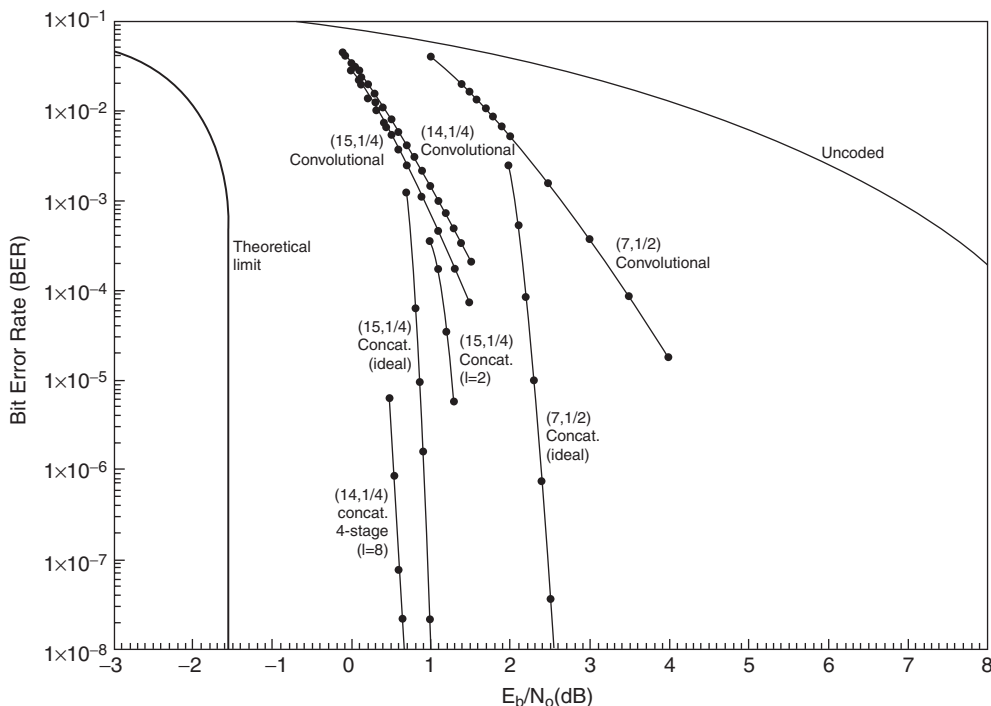


Figure 6. Telemetry communication channel performance for various coding schemes. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/return_5.htm.

added complexity of the codes, which has its greatest effect in the size of the ground decoder, clearly provides increased reliability in correct communication.

Research on new and even more powerful coding schemes such as turbo codes continued to occupy an important place in the Advanced Systems Program. Turbo codes are composite codes made up of short constraint length convolutional codes and a data stream interleaver (24,25). The decoding likewise consists of decoders for the simple component codes, but use iterative sharing of information between them. These codes that push hard on the fundamental theoretical limits of signal detection can result in almost a full decibel of performance gain over the best previous concatenated coding systems.

Data Compression. Source encoding and data compression are not typically considered part of the DSN's downlink functions, but the mathematics that underlie coding and decoding are a counterpart of those that guide the development of data compression. Simply stated, channel encoding is the insertion of structured redundancy into a data stream, whereas data compression is finding and removing of intrinsic redundancy. Imaging data are often highly redundant and can be compressed by factors of at least 2, and often 4 or more, without loss in quality. For Voyager, the combined effect of a very simplified image compression process, constrained to fit into available onboard memory, and the corresponding changes to the channel coding, was about a factor-of-2 increase in the number of images returned from Uranus and Neptune.

The success of data compression technology in enhancing the data return from the Voyager missions firmly established the technique as an important consideration in designing all future planetary downlinks. The original telecommunication link design for the Galileo spacecraft used data compression to double almost the amount of imaging data that the spacecraft could transmit from its orbital mission around Jupiter. The failure of the spacecraft's high-gain antenna before Galileo's arrival at Jupiter prompted an intense effort to find even more complex data compression schemes that would recover some of the Jupiter imaging data that otherwise could not have been returned (26).

Arraying of Antennae

The technique of antenna arraying, as practiced in the Deep Space Network, used the physical fact that a weak radio signal from a distant spacecraft received simultaneously by several antennae at different locations is degraded by a component of radio noise that is independent at each receiving station. By contrast, the transmitted spacecraft signal is dependent, or coherent, at each receiving site. In theory, therefore, the power of the signal, relative to the power of the noise, or signal-to-noise ratio, (SNR) can be improved by combining the individual antennae so that coherent spacecraft signals are reinforced and independent or noncoherent noise components are averaged.

In practice, this involved a complex digital process to compensate for the time, or phase, delays caused by the different distances between each station and the spacecraft and by the different distances between the various antenna locations and the reference station, where the combining function was carried out. This technique became known as antenna arraying, and the digital processing function that realized the theoretical "gain" of the entire process was called "signal combining" (27).

By 1970, conceptual studies had described and analyzed the performance of several levels of signal combining, and two of these schemes, carrier and baseband combining, were of potential interest to the Network. Both techniques involved compensation for the time, or phase, delays due to the various locations of the arrayed antennas. The difference lay in the frequency at which the combining function was performed. "Carrier" combining was carried out at the carrier frequency of the received signal, whereas "baseband" combining was carried out at the frequencies of the subcarrier and data signal that modulated it. Each had its advantages and disadvantages, but "baseband" combining proved easier to implement and was, obviously, tried first.

The "arraying and signal combining" concept was first developed and demonstrated in 1969 and 1970 by J. Urech, a Spanish engineer working at the Madrid tracking station (28,29). Using signals from the Pioneer 8 spacecraft and a microwave link to connect two 26-m stations located 20 m apart (DSS 61 and DSS 62), he succeeded in demonstrating for the first time the practical application of the principle of baseband combining in the Network. Because of the low baseband frequency of the Pioneer 8 data stream (8 bits per second) and close proximity of the antennae, no compensation for time delay was necessary.

Within the bounds of experimental error, this demonstration confirmed the R&D theoretical estimates of performance gain and encouraged the JPL researchers to press forward with a more complex form of baseband combining at a much higher data rate (117 kilobits per second) in real time.

The demonstration took place at Goldstone in September 1974, using the downlink signals from the Mariner–Venus–Mercury (MVM) spacecraft during its second encounter with the planet Mercury (30). Spacecraft signals from the two 26-m antennae, DSS 12 and DSS 13, were combined in an R&D combiner with signals from the DSS 14 64-m antenna in real time at 117 kbps. The less-than-predicted arraying gain obtained in this demonstration (9 versus 17%) was attributed to small differences in performance between key elements of the several data processing systems involved in the test. Although this experience demonstrated the practical difficulty of achieving the full theoretical gain of an arrayed antenna system and the critical effect of very small variations in the performance of its components, it also established the technical feasibility of baseband arraying of very weak high-rate signals.

In 1977, using the lessons learned from these demonstrations as background, the DSN started to develop an operational arraying capability for the Network. The Voyager 1 and 2 encounters with Saturn in 1980 and 1981, respectively, would be the first to use the arraying in the Network. A prototype baseband real-time combiner (RTC), based on the analysis and design techniques developed by the earlier R&D activity, was completed in the fall of 1978. Designed to combine the signals from DSS 12 and DSS 14 at Goldstone, it was used with varying degrees of success to enhance the signals from the Voyagers at Jupiter in March and July 1979 and the Pioneer 11 encounter with Saturn in August and September of that year.

Like the previous demonstration, this experience emphasized the critical importance of having all elements of the array—receivers, antennae, and instrumentation—operating precisely according to their specified performance capabilities. With this very much in mind, the DSN proceeded to the design of operational versions of the RTC for use at all three complexes to support the Voyager 1 and 2 encounters with Saturn. The operational versions of the RTC embodied many improvements derived from the experience with the R&D prototype. By mid-August 1980, they were installed and being used to array the 64-m and 34-m antennae at all three complexes, as Voyager 1 began its far-encounter operations. During this period, the average arraying gain was 0.62 dB, that is, about 15% greater than that of the 64-m antenna alone. While this was “good,” improvement came slowly as more rigorous control and calibration measures for the array elements were instituted throughout the Network. By the time Voyager 2 reached Saturn in August 1981, these measures, supplemented by additional training and calibration procedures, had paid off. The average arraying gain around the Network increased to 0.8dB (approximately 20%), relative to the 64-m antenna alone, clearly a most satisfactory result and the best up to that time. Antenna arraying had become a permanent addition to the capability of the Network.

While researchers working within the Advanced Systems Program continued to explore new processes for arraying antennas, engineers within the DSN took advantage of the long flight time between the Voyager Saturn and

Uranus encounters to refine the existing RTC configuration. During the next 5 years, the formerly separate data processing functions of combining, demodulation, and synchronization were integrated into a single assembly, and performance, stability, and operational convenience were improved. By the time Voyager 2 approached Uranus in 1985, the new baseband assemblies (BBAs), as they were called, had been installed at all three complexes. In addition, a special version of the basic four-antenna BBA was installed at the Canberra Complex. This provided for combining the Canberra array of one 64-m and two 34-m antennae with signals from the 64-m Parkes Radio Telescope, 200 km distant (Fig. 7).

In January 1986, this arrangement was a key factor in the successful return of Voyager imaging data from the unprecedented range of Uranus (20 AU). But even greater achievements in antenna arraying lay ahead.

In 1989, the DSN used a similar arrangement with great success to capture the Voyager imaging data at a still greater range—from Neptune (30 AU). This time the Goldstone 70-m and 34-m antennas were arrayed with the 27 antennas of the Very Large Array (VLA) (Fig. 8) of the National Radio Astronomy Observatory at Socorro, New Mexico (31).

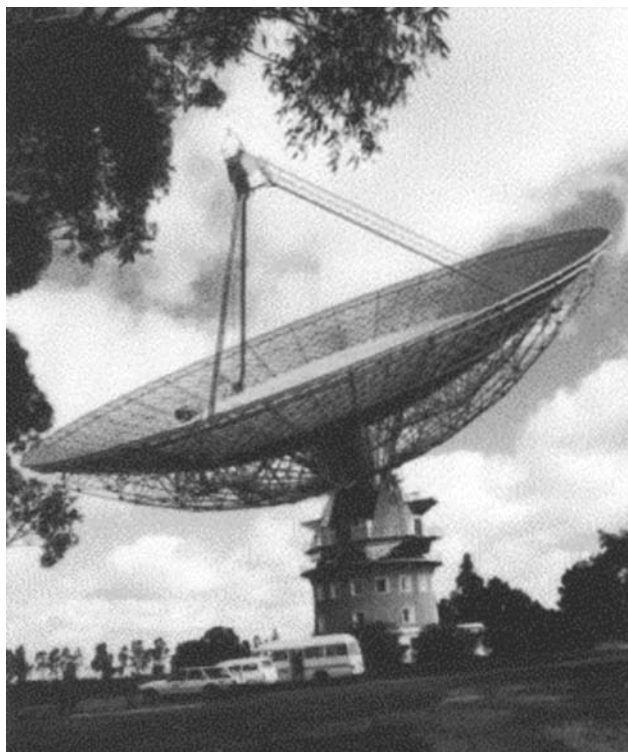


Figure 7. Parkes Radio Telescope, Australia. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/array_1.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



Figure 8. The Very Large Array (VLA) of the National Radio Astronomy Observatory (NRAO) at Socorro, New Mexico. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/array_2.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

DSN support for the Voyager encounter with Uranus was further augmented by the Canberra–Parkes array in Australia, which the DSN had reinstated with the addition of new BBAs, new 34-m antennae, and the upgraded 70-m antenna.

The success of these applications of the multiple antenna arraying technique provided the DSN with a solid background of operational experience. The DSN drew heavily on this experience a few years later when it was called upon to recover the scientific data from Galileo after the failure of the spacecraft's high-gain antenna in 1991. Together with the data compression and coding techniques discussed earlier, the Network's Canberra/Parkes/Goldstone antenna arrays succeeded in recovering a volume of data that, according to the Galileo project, was equivalent to about 70% of that of the original mission.

With time, arraying of multiple antennae within complexes, between complexes, or between international space agencies, came into general use to enhance the downlink capability of the Network. In the latter years of the century, most of the enhancements in arraying in the DSN were driven by implementation and operational considerations, rather than by new technology, although the Advanced Systems Program continued to explore the boundaries of performance of various alternative arraying architectures and combining techniques.

Radio-Metric Techniques

In addition to being able to exchange forward and return link data with an exploring spacecraft, it is equally important to know the precise location of the spacecraft, its speed, and direction (velocity). Information about the position and velocity of the spacecraft can be extracted from the one-way or two-way radio signals passing between the spacecraft and the DSN. When these data are extracted by appropriate processing and further refined to remove aberrations

introduced by the propagative medium along the radio path between spacecraft and Earth, it can be used for spacecraft "navigation."

Radio-metric techniques similar to those used for spacecraft navigation can also be used for more explicit scientific purposes, notably radio science, radio astronomy, and radio interferometry on very long baselines (VLBI).

Since its inception, the DSN Advanced Systems Program has worked to develop effective radio-metric tools, techniques, observing strategies, and analytical techniques that furthered the DSN preeminence in these unique fields of science. More recently, the program demonstrated the application of Global Positioning System (GPS) technology to refine further radio-metric data generated by the DSN.

Doppler and Range Data. If Earth and the spacecraft were standing still, the time for a radio signal to travel from Earth to the spacecraft and back would be a measurement of the distance between them. This is referred to as the round-trip light time (RTLTL). However, because Earth and the spacecraft are both in motion, the RTLTL contains both position and velocity information, which can be disentangled through multiple measurements and suitable analysis. The precision of such measurements is limited by the precision at which one can attach a time-tag marker to the radio signals and by the strength of the signal in proportion to the noise mixed with it, or by the signal-to-noise ratio (SNR).

Precise measurements of changes in this light time are far easier to obtain by observing the Doppler effect resulting from the relative motions. Such measurements are mechanized via the phase-locked loops in both spacecraft and ground receivers using the spacecraft's replica of the forward link residual carrier signal to generate the return link signal and counting the local replica of the return link residual carrier against the original carrier for the forward link signal. The raw precision of these measurements is comparable to the wavelength of the residual carrier signal, that is, a few centimeters for an X-band signal (8 GHz). Numerous error sources tend to corrupt the accuracy of the measurement and the inferred position and velocity of the spacecraft derived therefrom. The observed Doppler contains numerous distinct contributions, including the very significant component due to the rotation of Earth. As Earth turns, the position of any specific site on the surface describes a circle, centered at the spin axis of Earth, falling in a plane defined by the latitude of that site. The resultant Doppler component varies diurnally with a sinusoidal variation, which is at its maximum positive value when the spacecraft is first observable over the eastern horizon, and is at its corresponding negative value as it approaches the western horizon. A full-pass Doppler observation from horizon to horizon can be analyzed to extract the apparent spacecraft position in the sky, although the determination is somewhat weak near the equatorial plane. Direct measurements of the RTLTL are useful for resolving this difficulty (32).

Three distinct generations of instruments, designed to measure the RTLTL, were developed by the Advanced Systems Program and used in an ad hoc fashion for spacecraft support before a hybrid version was designed and implemented around the DSN (33–35). The third instrument designed, the Mu-II Ranging Machine, was used with Viking Landers in a celestial mechanics experiment, which provided the most precise test, until that time, of the general theory of relativity (36).

These devices function by imposing an additional “ranging” modulation signal on the forward link, which is copied on the spacecraft (within the limits imposed by noise) and then imposed on the return link. The ranging signal is actually a very long-period coded sequence that provides the effect of a discrete time tag. The bandwidth of the signal is of the order of 1 MHz, giving the measurement a raw precision of a few hundred meters, resolvable with care to a few meters. Among other features, the Mu-II Ranging Machine included the first demonstrated application of digital detection techniques that would figure strongly in future developments for the DSN.

Timing Standards. The basic units of measurement for all radio-metric observations, Doppler or range, derive from the wavelength of the transmitted signal. Uncertainties or errors in the knowledge of that wavelength are equivalent to errors in the derived spacecraft position. The need for accurate radio metrics has motivated the DSN to develop some of the most precise, most stable frequency standards in the world. Although the current suite of hydrogen maser frequency standards in DSN field sites was built outside of JPL, the design is the end product of a long collaboration in technology development; research units were built at JPL and elsewhere (37,38).

Continued research for improved frequency standards resulted in the development of a new linear ion trap standard (Fig. 9) that offered improved long-term stability of a few parts in 10^{16} , as well as simpler and easier maintenance than that required by hydrogen masers (39).

Earth’s Rotation and Propagative Media. Radio-metric Doppler and range data enable determining the apparent location of a spacecraft relative to the position and attitude of rotating Earth. Earth, however, is not a perfectly rigid body at constant rotation, but contains fluid components as well, which slosh about and induce variations in rotation of perhaps a few milliseconds per day.

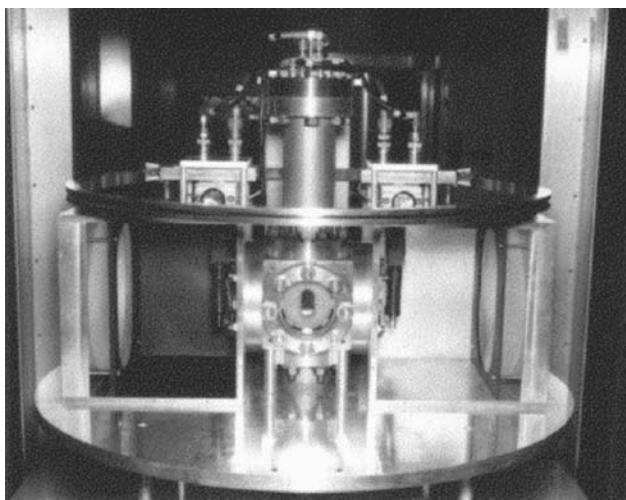


Figure 9. New linear ion trap frequency standard. This figure can also be seen at the following website: http://deepspace.jpl.nasa.gov/technology/95_20/radio_1.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Calibration of Earth's attitude is necessary so that the spacecraft's position in inertial space can be determined, which is a necessary factor in navigating it toward a target planet. Such calibration is available via the world's optical observatories, and with greater precision via radio techniques, which will be discussed further in later sections titled "VLBI and Radio Astronomy" and "Global Positioning System."

The interplanetary media along the signal path between Earth and the spacecraft affect the accuracy of Doppler and range observations. The charged ions in the tenuous plasma that spreads out from the Sun, known as the solar wind, slightly bend and delay the radio signal. Likewise, charged ions in Earth's own ionosphere and water vapor and other gases of the denser lower atmosphere bend and delay the radio signal. All of these factors are highly variable because of other factors, such as the intensity of solar activity, season, time of day, and weather. All factors must be calibrated, modeled, or measured to achieve the needed accuracy. Over the years, the DSN Advanced Systems Program has devised an increasingly accurate series of tools and techniques for these calibrations (37–43).

Radio Science. Radio science is the term used to describe the scientific information obtained from the intervening pathway between Earth and a spacecraft by using radio links. The effects of the solar wind on the radio signal path interfere with our efforts to determine the location of a spacecraft, but if the relative motions of Earth and a spacecraft are modeled and removed from the radio-metric data, much of what remains is information about the solar wind and, thus, about the Sun itself. Other interfering factors are also of scientific interest.

In some situations, the signal path passes close by a planet or other object, and the signal itself is bent, delayed, obscured, or reflected by that object and its surrounding atmosphere. These situations provide a unique opportunity for scientists to extract information from the signal about object size, atmospheric density profiles, and other factors not otherwise observable. Algorithms and other tools devised to help calibrate and remove interfering signatures from radio-metric data for use in locating a spacecraft often become part of the process for extracting scientific information from the same radio-metric data stream. The precision frequency standards, low-noise amplifiers, and other elements of the DSN are key factors in the ability to extract this information with scientifically interesting accuracy. Occasionally, engineering model equipment is placed in the Network in parallel with operational instrumentation for ad hoc support of metric data gathering for some unique scientific event.

The effects of gravity can also be observed by means of the radio link. Several situations are of interest. If the spacecraft is passing by or in orbit about an object that has a lumpy uneven density, that unevenness will cause a variation in the spacecraft's pathway that will be observable via radio-metric data. If the radio signal passes near a massive object such as the Sun, the radio signal's path will be bent by the intense gravity field, according to the theories of general relativity. And in concept, gravitational waves (a yet-to-be observed aspect of gravity field theory) should be observable in the Doppler data from a distant spacecraft (44).

VLBI and Radio Astronomy. The technical excellence of the current DSN is, at least in part, a result of long and fruitful collaboration with an active radio

astronomy community at the California Institute of Technology (Caltech) and elsewhere. Many distant stars, galaxies, and quasars are detectable by the DSN at radio frequencies. The furthest of them are virtually motionless when viewed from Earth and can be considered a fixed-coordinate system to which spacecraft and other observations can be referenced. Observations relative to this coordinate set help to reduce the distorting effects of intervening material in the radio signal path and uncertainties in the exact rotational attitude of Earth during spacecraft observations.

Little precise information can be extracted by observing these objects one at a time and from a single site, but concurrent observation at a pair of sites will determine the relative position of the two sites referenced to the distant object. The observing technique is known as very long baseline interferometry (VLBI). If three sites are used in VLBI pairs and multiple objects are observed, the positional attitude of Earth and the relative positions of the observed objects can be determined. If one of the observed items is a spacecraft transmitting a suitable signal, its position and velocity in the sky can be very accurately defined (45). A demonstration of this technique led to operational use for spacecraft such as Voyager and Magellan.

VLBI can also be used in conjunction with conventional radio-metric data types to provide calibration for the positional attitude of Earth. Such observations can be made without interfering with spacecraft communication, except for the time use of the DSN antennae. In addition to determining the attitude of Earth, the observations measure the relative behavior of the frequency standards at the widely separated DSN sites, and thus help to maintain their precise performance.

The DSN equipment and software needed for VLBI signal acquisition and signal processing (correlation) was designed and developed in a collaboration involving the Advanced Systems Program, the operational DSN, and the Caltech radio astronomy community. The tools needed to produce VLBI metric observations for the DSN were essentially the same as those for interferometric radio astronomy. Caltech was funded by the National Science Foundation for this activity. Both Caltech and the DSN shared in the efforts of the design and obtained products that were substantially better than any that they could have been obtained independently.

Another area of common interest between the DSN and the radio astronomy community is that of precision wideband spectral analysis. Development efforts produced spectral analytical tools that have been employed by the DSN in spacecraft emergencies and in examining the DSN's radio interference environment and have served as preprototype models for equipment for the DSN (46,47). Demonstration of the technical feasibility of very wide band spectral analysis and preliminary observations by a mega channel spectrum analyzer fielded by the Advanced Systems Program helped establish the sky survey planned as part of the former SETI (Search for Extraterrestrial Intelligence) Program.

Another technique (one similar to using of VLBI as a radio-metric reference) is used if two spacecraft are flown to the same target. The second can be observed relative to the first and can provide better target-relative guidance, once the first has arrived at the target.

The Global Positioning System. The Global Positioning System (GPS) is a constellation of Earth-orbiting satellites designed (initially) to provide for military navigation on Earth's surface. Research under the Advanced Systems Program showed that these satellites could provide an excellent tool to calibrate and assist in radio-metric observations of distant spacecraft. GPS satellites fly above Earth's atmosphere and ionosphere in well-defined orbits, so that their signals can be used to measure the delay through these media in a number of directions. Using suitable modeling and analysis, these measurements can be used to develop the atmospheric and ionospheric calibrations for the radio path to a distant spacecraft (48).

Additionally, because GPS satellites are in free orbit about Earth, their positions are defined relative to the center of mass of Earth, not its surface. They provide another method for observing the uneven rotation of Earth.

GPS techniques can also be used to determine the position of an Earth-orbiting spacecraft relative to the GPS satellites, as long as the spacecraft carries a receiver for the GPS signals. GPS signals were subsequently used by the TOPEX/POSEIDON Project for precise orbit determination and consequent enhancement of its scientific return (49).

Goldstone Solar System Radar

The Goldstone Solar System Radar (GSSR) is a unique scientific instrument for observing nearby asteroids, the surfaces of Venus and Mars, the satellites of Jupiter, and other objects in the solar system. Although the GSSR uses the DSN 70-m antenna for its scheduled observing sessions, its receiving, transmitting, and data processing equipment is unique to the radar program. The GSSR is a product of many years of development. In the early days of the DSN, the Advanced Systems Program took ownership of the radar capability at the DSN's Goldstone, California, site and evolved and nurtured it as a vehicle for developing and demonstrating many of the RF and signal processing capabilities that the Network would need elsewhere.

Scientific results abounded as well, but were not its primary product. Timely development of DSN capabilities was the major result. Preparations for radar observations of asteroids at the DSN Technology Development Field Site bore many resemblances to those for a spacecraft planetary encounter, because the radar observations could be successfully accomplished only during the few days when Earth and the radar target were closest together.

In the conventional formulation of the radar sensitivity equations, that sensitivity depends on the aperture, temperature, power, and gain of the system elements. Here, aperture refers to the effective size, or collecting area of the receiving antenna, and temperature is a way of referring to the noise in the receiving system, a lower temperature means less noise. Power refers to the raw power level from the transmitter, and gain is the effective gain of the transmitting antenna, which depends in turn upon its size, its surface efficiency, and the frequency of the transmitted signal. When the same antenna is used both to transmit and receive, the antenna size and efficiency appear twice in the radar equations.

Significant improvements in the DSN's capability for telemetry reception were to come from the move upward in frequency from S band (2 GHz) to X band

(8 GHz) on the large 64-m antennae. Performance of these antennae at higher frequencies and the ability to point them successfully were uncertain, however, and these uncertainties would best be removed by radar observations before spacecraft with X-band capabilities were launched (50). The radar had obvious benefit from the large antenna and the higher frequency. The first flight experiment for X-band communication was carried out on the 1973 Mariner Venus Mars mission. Successful radar observations from the Goldstone 64-m antenna demonstrated that the challenge of operating the large antennae at the higher X-band frequency could be surmounted.

High-power transmitters were needed by the DSN for its emergency forward-link functions, but were plagued by problems such as arcing in the waveguide path when power densities became too high. High-power transmitters were essential for the radar to “see” at increased distances and with increased resolution. Intense development efforts at the DSN Technology Development Field Site could take place without interference or risk to spacecraft support in the Network. Successful resolution of the high-power problems for the radar under the Advanced Systems Program enabled successful implementation of the high-power capability needed by the Network for uplink communications.

Low-noise amplifiers were needed by the DSN to increase data return from distant spacecraft. Low-noise amplifiers are also essential for the radar to enable it to detect echoes from increasingly distant targets or to provide increased resolution of already detectable targets. The synergistic needs of both the radar system and the Network led to development of the extremely low-noise maser amplifiers that became part of the standard operational inventory of the DSN.

Digital systems technology was rapidly evolving during this period and would play an increasing role in the developing DSN. Equipment developed by the Advanced Systems Program for its radar application included (1) digital encoders to provide spatial resolution of parts of the radar echo, (2) computer-driven programmable oscillators to accommodate Doppler effects on the signal path from Earth-to-target-to-Earth, and (3) complex high-speed digital signal processing and spectral analysis equipment. Much of the digital technology developed this way would transfer quickly to other parts of the signal processing work under the Advanced Systems Program and eventually into the operational DSN. Some of the elements would find direct application, such as programmable oscillators, which became essential for maintaining contact with the Voyager 2 spacecraft following a partial failure in its receiver soon after launch. And the signal analysis tools developed to understand and optimize radar would be called on many times over the years to help respond to spacecraft emergencies.

Some of the products of early radar observations were both scientific in nature and essential for providing information for planning and executing of NASA's missions. One notable “first” was direct measurement of the astronomical unit (51). (One astronomical unit (AU) is equal to 1.5×10^8 km, i.e., the mean distance between Earth and the Sun.) It sets the scale size for describing distances in the solar system. The measurement was made to support preparations for Mariner 2 to Venus and provided a correction of 66,000 km from conventional belief at that time. It also enabled corrections that brought the mission into the desired trajectory for its close flyby of the planet. The GSSR was also used in qualifying potential Mars landing sites for the Viking Landers and continues to

provide information about the position and motion of the planets, which is used to update the predicted orbits for the planets of the solar system.

Telecommunications Performance of the Network

The progress of deep space communications capability during the 40-year history of the Network is shown in Fig. 10. In Fig. 10, the timeline on the horizontal axis of the figure covers the first 40 years of actual Network operational experience through the close of the century, and extends for a further 20 years to forecast the potential for future improvements through the year 2020. The vertical axis displays the growth in space-to-Earth, downlink capability of the Network. The downlink capability is given in units of the telemetry data rate (bits per second) on a logarithmic scale, and represents the equivalent data rate capability for a typical spacecraft at Jupiter distance (750 million kilometers). Significant events in the history of deep space telecommunications and deep space exploration are appropriately annotated on both axes.

In interpreting the data presented in Fig. 10, it will be observed that the logarithmic scale that displays the data rate gives an impression that the early improvements are more significant than the later improvements because the steps represent fractional or percentage increases, rather than the magnitudes of the actual increases, which are much larger in the later years.

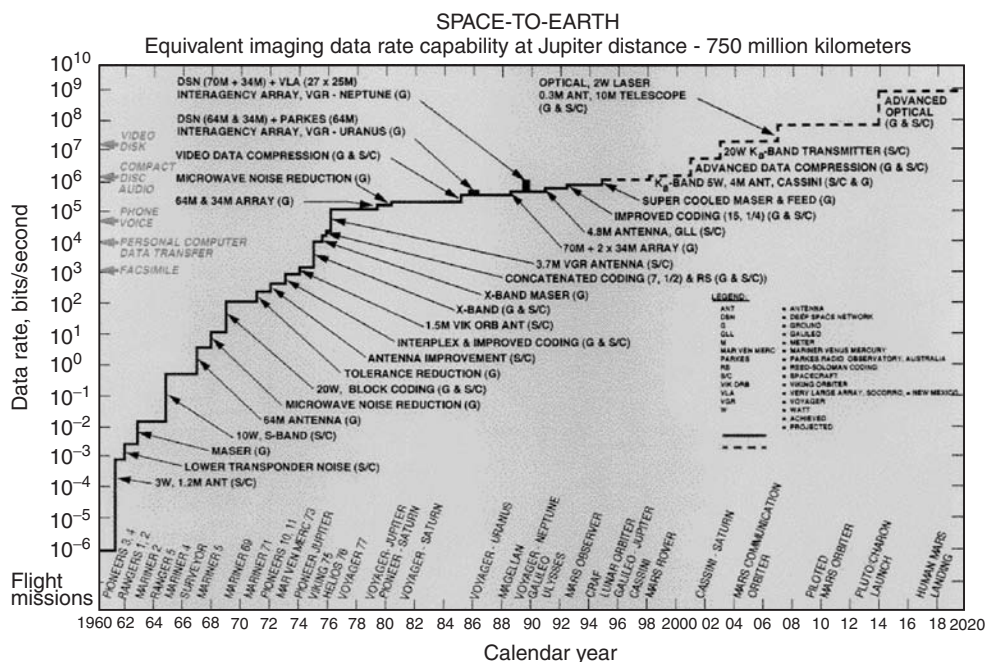


Figure 10. Profile of deep space telecommunications performance: 1960–2020. This figure can also be seen at the following website: <http://deepspace.jpl.nasa.gov/dsn/history/album/images/dsn71.gif>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Presented this way, however, the figure clearly shows that, from inception through 1997, the downlink capability of the Network grew from an equivalent Jupiter data rate of 10^{-6} bits/s to nearly 10^{+6} bits/s. Note, however, that the drive to switch to “faster-better-cheaper” missions in the late 1990s resulted in decreased capabilities on the spacecraft and eroded the actually implemented capabilities by approximately two orders of magnitude. Nevertheless, that still represents 10 orders of magnitude of improvement.

This remarkable progress is not, of course, solely due to improvements in the Network. Many of the steps result from “cooperative” changes on the part of both the DSN and the spacecraft. Coding, for example, is applied to the data on the spacecraft and removed on Earth. A change in frequency has resulted in some of the larger steps shown by making the radio beam from the spacecraft more narrowly focused. Such a change necessitates equipment changes on both the spacecraft and on Earth.

Other steps represent advances that are strictly spacecraft related, such as increases in return-link transmitter power or increases in spacecraft antenna size, which improves performance by more narrowly focusing the radio beam from the spacecraft. Still other steps depict improvements strictly resulting from the DSN, such as reducing receiving system temperature, increasing the size of ground antennae, or using arrays of antennae to increase the effective surface area available for collecting signal power.

Other Deep Space Network Activities

A number of other DSN activities and developments are worth mentioning.

Cost Reduction Initiatives. Pressures have always existed to reduce costs in the DSN. In 1994, a Network automation work area was set up to develop automated procedures that would replace the extremely operator-intensive work of running a DSN tracking station during a spacecraft “pass.” This effort soon produced demonstrable results. A fully automated satellite-tracking terminal for near-Earth satellites was demonstrated in 1994 (52). A software prototype controller that reduced the number of manual inputs for a typical 8-hour track from 900 to 3 was installed at DSS 13 and used to support Ka-band operations at that site (53). Eventually, this technology found its way into the operational Network. A contract for a new, small, deep space transponder offering lower size and power needs, and most importantly lower production costs, was initiated with Motorola in July 1995 (54). This became an element in JPL’s future low-cost (faster-better-cheaper) spacecraft.

Prototypes of a new class of low-cost, fully automated, autonomous ground stations that would simplify implementation and operation and reduce the life-cycle cost of tracking stations in the DSN were introduced in 1995, 1996, and 1997. The first of these terminals was designed for tracking spacecraft in low Earth orbit and was named LEO-T. It was enclosed in a radome and mounted on the roof of a building at JPL where it accumulated more than 2 years of unattended satellite tracking operations without problems. Prompted by the success of LEO-T, the program undertook a fast-track effort to develop a similar automated terminal for deep space applications. It would be called DS-T. The

prototype DS-T was to be implemented at the 26-m BWG antenna at Goldstone. Automation technology was carried one step further into the area of Network operations in 1997, when Automated Real-time Spacecraft Navigation (ARTSN) was introduced (55). In addition to the antenna system, the DS-T included an X-band microwave system, a 4-kW transmitter, and an electronics rack containing commercial, off-the-shelf equipment to carry out all baseband telemetry downlink, command uplink, and Doppler and ranging functions. It was planned to demonstrate DS-T with the Mars Global Surveyor early in 1998 and to use this technology (autonomous uplink and downlink) in the Network with the New Millennium DS-1 spacecraft later that year.

Ka-Band Development. In the past, the most significant improvements in the Network's communication capabilities were made by moving to higher frequency bands. Recognizing this, research and development in the Ka band (32 GHz) was started in 1980. Initial efforts were directed toward low-noise amplifier development and system benefit studies. However, it was also clear that the performance of existing antennas (which were designed for much lower frequencies) would severely limit the improvement in performance that could be realized from the higher operating frequency. Accordingly, in 1991, a new antenna specifically designed for research and development in the Ka band was installed at the DSS 13 Venus site at Goldstone. It was used as the pathfinder for developing large-aperture beam waveguide (BWG) antennae that were, in due course, implemented throughout the Network.

Small imperfections in the surface of an antenna cause larger degradations in the Ka band than at lower frequencies, and, because of the narrower beam width, small pointing errors have a much larger effect. In 1994, improvements in antenna efficiency and in antenna pointing were made in the research and development Ka-band antenna at DSS 13 and in the new operational antenna at DSS 24 (56). These improvements were effected by using microwave holography for precise determination of antenna efficiency and a special gravity compensation system to counteract the effect of gravitational sag as a function of antenna elevation angle. In a search for further downlink improvement, a new feed system consisting of a maximally compact array of seven, circular, Ka-band horns was designed and tested at DSS 13 (57). Each horn was connected to a cryogenically cooled low-noise amplifier, a frequency down-converter, an analog to digital converter, and a digital signal processor. The signals from each horn were optimally combined in a signal processor and presented as a single output whose quality was equivalent to that of a signal from an undistorted antenna. The measured gain in downlink performance was 0.7 dB.

The technology program continued to develop operational concepts that eventually led to the adoption of the Ka band for deep space missions. Trade-off studies between the X band and Ka band showed that the overall advantage of the Ka band, taking due account of the negative effects inherent in its use, was about a factor of 4, or 6 dB, in data return capability (58). Obviously, end-to-end system demonstrations were needed to instill confidence in the new technology. In 1993, the Mars Observer spacecraft carried a non linear element in its transmission feed to produce a fourth harmonic of the X-band signal. The demonstration (called KABLE for KA-band link experiment) provided a weak Ka-band signal for the tracking antenna at DSS 13. A second demonstration (KABLE II)

was conducted using the Mars Global Surveyor mission in 1996. This experiment was used to characterize Ka-band link performance under real flight conditions and to validate the theoretical models derived from the studies mentioned above.

KABLE II required additions to the MGS spacecraft radio transponder to generate a modulated Ka-band downlink from which the improvements that had been made to the DSS 13 to improve Ka-band performance could be evaluated. Though the main objective of KABLE II was to evaluate the Ka band for future operational use, it also served as a test bed for new Ka-band technology applications in both flight systems and ground systems.

In the course of transition to simultaneous X/Ka-band operation, the DSN needed the capability to support various combinations of X- and Ka-band uplinks and downlinks. These included Ka-band receive only, X/Ka-band simultaneous receive, with or without X-band transmit, and full X-band transmit/receive simultaneously with Ka-band receive/transmit. The technology program developed new microwave techniques using frequency selective surfaces and feed junction diplexers to provide the frequency and power isolation necessary to realize the performance required by a practical device. In late 1996, a demonstration at DSS 13 showed that these four different modes of operation could coexist on a single, beam waveguide antenna within acceptable performance limits (59). This work provided a viable solution to the problem of simultaneous X/Ka-band operation on a single antenna that would be needed by the operational network to support the Cassini radio science experiment (search for gravitational waves) in 2000.

Recognizing the need for a cheaper, smaller, less power-consuming radio transponder to replace the existing device on future deep space missions, the Advanced Development Program embarked on a joint program with other JPL organizations to develop the Small Deep Space Transponder (SDST). The concept employed microwave monolithic integrated circuits in the RF circuits and application specific integrated circuit (ASIC) techniques to perform digital signal processing functions and a RISC microprocessor to orchestrate overall transponder operation (60). The transponder would transmit coherent X-band and Ka-band downlinks and receive an X-band uplink. Besides minimizing production costs, the principal design drivers were reductions in mass, power consumption, and volume. The Small Deep Space Transponder was flown in space for the first time onboard the DS-1 New Millennium spacecraft in July 1998.

To provide better understanding of the performance of Ka-band links relative to X-band links from the vantage point of a spaceborne radio source, the DSN technology program engaged in developing a small, low Earth orbiting spacecraft called SURFSAT-1. Launched in 1995, the experiment provided an end-to-end test of Ka-band signals under all weather conditions and DSS 13 antenna elevation angles, as the spacecraft passed over Goldstone. The SURFSAT data was also used for comparison with the KABLE data received from MGS. Later, the SURFSAT X-band and Ka-band downlinks were used to great advantage to test and calibrate the DSN's new 11-meter antennas, before their support of the VSOP (HALCA) Orbiting VLBI mission in 1996.

Optical Communications Development. Beginning in 1980, the technology program supported theoretical analyses that predicted, under certain system

and background light conditions typical of deep space applications, the ability to communicate at more than 2.5 bits of information per detected photon at the receiver. Laboratory tests later confirmed these theoretical predictions (61,62). However, detection power efficiency was only one of the many factors that needed to be studied to bring optical communications to reality. Others included laser transmitter efficiency, spatial beam acquisition, tracking and pointing, link performance tools, flight terminal systems design, definition of cost-effective ground stations, and mitigation of Earth's atmospheric effects on ground stations.

Several system-level demonstrations were carried out as this work progressed. The first, in December 1992, involved detecting a ground-based pulsed laser transmission by the Galileo spacecraft during its second Earth fly by. Transmitted laser signals generated from two ground-based telescope facilities were successfully detected at spacecraft–Earth distances up to 6 million km (63). The second demonstration was carried out during the period November 1995 to May 1996 between JPL's optical telescope facility at Table Mountain, California, and the Japanese Earth-orbiting satellite ETS VI at geosynchronous Earth-orbit altitudes (40,000 km). Data-modulated transmissions were successfully detected in both uplink and downlink directions at a data rate of 1 Mbps (64).

Both experiments yielded important observational data in support of theoretical studies and encouraged the further development of optical communications technology with follow-on supportive flight demonstrations.

DSN Science. Science and technology were always closely coupled in the DSN. Since the very beginnings of the DSN, its radio telescopes provided world-class instruments for radio astronomy, planetary radar, and radio science. Many technology program achievements were of direct benefit to these scientific endeavors, and DSN science activities frequently resulted in new techniques that eventually found their way into the operational Network.

In the period reported here, the program supported radio astronomy investigations related to the formation of stars, to the study of microwave radio emissions from Jupiter, and to radio science measurements of the electron density in the solar plasma outside the plane of the ecliptic (Ulysses spacecraft). The DSN also supported a program of tropospheric delay measurements, which would be of direct benefit to the Cassini gravitational wave experiment. The Goldstone Solar System Radar (GSSR) continued its highly successful series of Earth-crossing asteroid (ECA) observations, which began in 1992 with images of Toutatis (asteroid 4179) and Geographos (asteroid 1620) in 1994 and Golevka (asteroid 6489) in 1995 (65). This work was expected to increase in the years ahead as new and improved optical search programs enabled the discovery of more ECAs.

Note

Douglas J. Mudgway, currently of Sonoma, California, was formerly with JPL.

BIBLIOGRAPHY

1. Mudgway, D.J. Uplink–Downlink: A History of the Deep Space Network 1957–1997. National Aeronautics and Space Administration (NASA History Series). U.S. Government Printing Office, Washington, DC, 2001.

2. Layland, J.W., and L.L. Rauch. *The Evolution of Technology in the Deep Space Network, A History of the Advanced Systems Program*. Jet Propulsion Laboratory, Pasadena, CA, 1995.
3. Potter, P.D. Shaped Antenna Designs and Performance for 64-m Class DSN Antennas, DSN Progress Report 42-20. Jet Propulsion Laboratory, Pasadena, CA, 1974, pp. 92–111.
4. Fanelli, N.A., J.P. Goodwin, S.M. Petty, T. Hayashi, T. Nishimura, and T. Takano. Utilization of the Usuda Deep Space Center for the United States International Cometary Explorer (ICE). *Proc. 15th Int. Symp. Space Technol. Sci.*, Tokyo, Japan, 1986.
5. Benjauthrit, B., and T.K. Truong. Encoding and Decoding a Telecommunication Standard Command Code, DSN Progress Report 42–38. Jet Propulsion Laboratory, Pasadena, CA, 1977, pp. 115–119.
6. Clauss, R.C., and E. Wiebe. Low-Noise Receivers: Microwave Maser Development, JPL-Tech. Report 32-1526, Vol. XIX. Jet Propulsion Laboratory, Pasadena, CA, 1974, pp. 93–99.
7. Glass, G.W., G.G. Ortiz, and D.L. Johnson. X-Band Ultralow-Noise Maser Amplifier Performance, TDA Progress Report 42–116. Jet Propulsion Laboratory, Pasadena, CA, 1994, pp. 246–253.
8. Tanida, L. An 8.4-GHz Cryogenically Cooled HEMT Amplifier for DSS 13, TDA Progress Report 42-94. Jet Propulsion Laboratory, Pasadena, CA, 1988, pp. 163–169.
9. Ulvestad, J.S., G.M. Resch, and W.D. Brundage. X-Band System Performance of the Very Large Array, TDA Progress Report 42-92. Jet Propulsion Laboratory, Pasadena, CA, 1988, pp. 123–137.
10. Bautista, J.J., G.G. Ortiz, K.H.G. Duh, W.F. Kopp, P. Ho, P.C. Chao, M.Y. Kao, P.M. Smith, and J.M. Ballingall, 32-GHz Cryogenically Cooled HEMT Low-Noise Amplifiers,” TDA Progress Report 42-95. Jet Propulsion Laboratory, Pasadena, CA, 1988, pp. 71–81.
11. Jaffe, R., and E. Rechtin, Design and performance of phase-lock circuits capable of near-optimum performance over a wide range of input signal and noise level. *IRE Trans. Inf. Theory* IT-1: 66–76 (March 1955).
12. Lindsey, W.C. The Effect of RF Timing Noise in Two-Way Communications Systems. JPL-SPS 37-32, Vol. IV. Jet Propulsion Laboratory, Pasadena, CA, 1965, pp. 284–288.
13. Tausworthe, R.C. Theory and Practical Design of Phase-Locked Receivers, JPL Technical Report 32-819. Jet Propulsion Laboratory, Pasadena, CA, 1971.
14. Baumgartner, W.S, W. Frey, M.H. Brockman, R.W. Burt, J.W. Layland, G.M. Munson, N.A. Burow, L. Couvillon, A. Vaisnys, R.G. Petrie, C.T. Stelzried, and J.K. Woo. Multiple-Mission Telemetry System, JPL-SPS 37-46, Vol. III. Jet Propulsion Laboratory, Pasadena, CA, 1967, pp. 175–243.
15. Tausworthe, R.C., et al. High Rate Telemetry Project, JPL-SPS 37-54, Vol. II. Jet Propulsion Laboratory, Pasadena, CA, 1968, pp. 71–81.
16. Brown, D.H., and W.J. Hurd. DSN Advanced Receiver: Breadboard Description and Test Results, TDA-Progress Report 42-89. Jet Propulsion Laboratory, Pasadena, CA, 1987, pp. 48–66.
17. Hinedi, S. A Functional Description of the Advanced Receiver, TDA-Progress Report 42-100. Jet Propulsion Laboratory, Pasadena, CA, 1990, pp. 131–149.
18. Stiffler, J.J., and A.J. Viterbi. Performance of a Class of Q-Orthogonal Signals for Communication over the Gaussian Channel, JPL-SPS 37-32, Vol. IV. Jet Propulsion Laboratory, Pasadena, CA, 1965, pp. 277–281.
19. Green, R.R. A Serial Orthogonal Decoder, JPL-SPS 37-39, Vol. IV. Jet Propulsion Laboratory, Pasadena, CA, 1966, pp. 247–252.

20. Lushbaugh, W.A., and J.W. Layland. System Design of a Sequential Decoding Machine, JPL-SPS 37-50, Vol. II. Jet Propulsion Laboratory, Pasadena, CA, 1968, pp. 71–78.
21. Butman, S.A., L.J. Deutsch, and R.L. Miller. Performance of Concatenated Codes for Deep Space Missions, TDA-Progress Report 42-63. Jet Propulsion Laboratory, Pasadena, CA, 1981, pp. 33–39.
22. Divsalar, D., and J.H. Yuen. Performance of Concatenated Reed–Solomon/Viterbi Channel Coding, TDA-Progress Report 42-71. Jet Propulsion Laboratory, Pasadena, CA, 1982, pp. 81–94.
23. Dolinar, S.J. A New Code for Galileo, TDA-Progress Report 42-93. Jet Propulsion Laboratory, Pasadena, CA, 1988, pp. 83–96.
24. Divsalar, D., and F. Pollara. On the Design of Turbo Codes, TDA-Progress Report 42-123. Jet Propulsion Laboratory, Pasadena, CA, 1995, pp. 99–121.
25. Benedetto, S., D. Divsalar, et al. Serial concatenation of interleaved codes; performance analysis, design, and iterative decoding. *IEEE Trans. Inf. Theory*, 44: 909 (1998).
26. Ekroot, L., S.J. Dolinar, and K.M. Cheung. Integer Cosine Transform Compression for Galileo at Jupiter: A Preliminary Look, TDA-Progress Report 42-115. Jet Propulsion Laboratory, Pasadena, CA, 1994, pp. 110–123.
27. Wilck, H. A Signal Combiner for Antenna Arraying, DSN Progress Report 42-25. Jet Propulsion Laboratory, Pasadena, CA, 1975, pp. 111–117.
28. Urech, J.M. Telemetry Improvement Proposal for the 85-foot Antenna Network, Space Programs Summary 37-63, Vol. II. Jet Propulsion Laboratory, Pasadena, CA, 1970.
29. Urech, J.M. Processed Data Combination for Telemetry Improvement at DSS 62, JPL-Tech. Report 32-1526, Vol. II. Jet Propulsion Laboratory, Pasadena, CA, 1971, pp. 169–176.
30. Winkelstein, R.A. Analysis of the Signal Combiner for Multiple Antenna Arraying, DSN Progress Report 42-26. Jet Propulsion Laboratory, Pasadena, CA, 1975, pp. 102–118.
31. Brown, D.W., W.D. Brundage, J.S. Ulvestad, S.S. Kent, and K.P. Bartos. Interagency Telemetry Arraying for Voyager-Neptune Encounter, TDA-Progress Report 42-102. Jet Propulsion Laboratory, Pasadena, CA, 1990, pp. 91–118.
32. Hamilton, T.W., and W.G. Melbourne. Information Content of a Single Pass of Doppler Data from a Distant Spacecraft, JPL-SPS 37-39, Vol. III. Jet Propulsion Laboratory, Pasadena, CA, 1966, pp. 18–23.
33. Lushbaugh, W.L., and L.D. Rice. Mariner Venus 67 Ranging System Digital Processing Design, JPL-SPS 37-50, Vol. II. Jet Propulsion Laboratory, Pasadena, CA, 1968.
34. Goldstein, R.M. Ranging with Sequential Components, JPL-SPS 37-52, Vol. II. Jet Propulsion Laboratory, Pasadena, CA, 1968, pp. 46–49.
35. Martin, W.L. A Binary Coded Sequential Acquisition Ranging System, JPL-SPS 37-57, Vol. II. Jet Propulsion Laboratory, Pasadena, CA, 1969, pp. 72–81.
36. Shapiro, I.I., R.D. Reasenberg, P.E. Mac Neil, R.B. Goldstein, J. Brenckle, D.L. Cain, T. Komarek, A.I. Zygielbaum, W.F. Gudlihy, and W.H. Michael, Jr. The Viking relativity experiment. *J. Geophys. Res.* 82: 4329–4334 (1977).
37. Finnie, D. Frequency Generation and Control: Atomic Hydrogen Maser Frequency Standard, JPL-Tech. Report 32-1526, Vol. I. Jet Propulsion Laboratory, Pasadena, CA, 1971, pp. 73–75.
38. Dachel, P.R., S.M. Petty, R.F. Meyer, and R.L. Syndor. Hydrogen Maser Frequency Standards for the Deep Space Networks, DSN Progress Report 42-40. Jet Propulsion Laboratory, Pasadena, CA, 1977, pp. 76–83.
39. Prestage, J.D., G.J. Dick, and L. Maleki. New Ion Trap for Atomic Frequency Standard Applications, TDA Progress Report 42-97. Jet Propulsion Laboratory, Pasadena, CA, 1989, pp. 58–63.

40. Winn, F.B. Tropospheric Refraction Calibrations and Their Significance on Radio-Metric Doppler Reductions, JPL-TR 32-1526, Vol. VII. Jet Propulsion Laboratory, Pasadena, CA, 1972, pp. 68–73.
41. von Roos, O.H., and B.D. Mulhall. An Evaluation of Charged Particle Calibration by a Two-Way Dual-Frequency Technique and Alternatives to This Technique, JPL Tech. Report 32-1526, Vol. XI. Jet Propulsion Laboratory, Pasadena, CA, 1972, pp. 42–52.
42. Slobin, S.D., and P.D. Batelaan. DSN Water Vapor Radiometer Tropospheric Range Delay Calibration, DSN Progress Report 42-49, Jet Propulsion Laboratory, Pasadena, CA, 1979, pp. 136–145.
43. Roth, M., and T. Yunck. VLBI System for Weekly Measurement of UTI and Polar Motion: Preliminary Results, TDA Progress Report 42-58. Jet Propulsion Laboratory, Pasadena, CA, 1980, pp. 15–20.
44. Nelson, S.J., and J.W. Armstrong. Gravitational Wave Searches Using the DSN, TDA Progress Report 42-94. Jet Propulsion Laboratory, Pasadena, CA, 1988, pp. 75–85.
45. Miller, J.K. The Application of Differential VLBI to Planetary Approach Orbit Determination, DSN Progress Report 42-40, Jet Propulsion Laboratory, Pasadena, CA, 1977, pp. 84–90.
46. Satorius, E.H., M.J. Grimm, G.A. Zimmerman, and H.C. Wilck. Finite Wordlength Implementation of a Megachannel Digital Spectrum Analyzer, TDA Progress Report 42-86. Jet Propulsion Laboratory, Pasadena, CA, 1986, pp. 244–254.
47. Quirk, M.P. (Institute for Defense Analyses, New Jersey), H.C. Wilck, M.F. Garyantes, and M.J. Grimm. A Wideband, High-Resolution Spectrum Analyzer, TDA Progress Report 42-93. Jet Propulsion Laboratory, Pasadena, CA, 1988, pp. 188–198.
48. Lichten, S.M. Precise Estimation of Tropospheric Path Delays with GPS Techniques, TDA Progress Report 42-100. Jet Propulsion Laboratory, Pasadena, CA, 1990, pp. 1–12.
49. Guinn, J., J. Jee, P. Wolff, F. Lagattuta, T. Drain, and V. Sierra. TOPEX/POSEIDON Operational Orbit Determination Results Using Global Positioning Satellites, TDA Progress Report 42-116. Jet Propulsion Laboratory, Pasadena, CA, 1994, pp. 163–174.
50. Leu, R.L. X-Band Radar System, JPL-TR 32-1526, Vol. XIX. Jet Propulsion Laboratory, Pasadena, CA, 1974, pp. 77–81.
51. Carpenter, R.L., and R.M. Goldstein. Preliminary Results of the 1962 Radar Astronomy Study of Venus, JPL-SPS 37-20, Vol. IV. Jet Propulsion Laboratory, Pasadena, CA, 1963, pp. 182–184.
52. Golshan, N. DS-T development. *TMOD Technol. Program News*, Issue 8, Jet Propulsion Laboratory, Pasadena, CA, 1997.
53. Teitelbaum, L. Deep Space Station 13. *IPN-ISD Technol. Sci. News*, Issue 14, Jet Propulsion Laboratory, Pasadena, CA, 2001.
54. Zingales, S. Small deep space transponder development. *DSN Technol. Program News*, Issue 4, Jet Propulsion Laboratory, Pasadena, CA, 1995.
55. Cagahuala, L. ARTSN: An automated real-time spacecraft navigation system. *TMOD Technol. Sci. Program News*, Issue 9, Jet Propulsion Laboratory, Pasadena, CA, 1998.
56. Morabito, D.D. The Efficiency Characterization of the DSS 13, 34-meter Beam Waveguide Antenna at Ka-band (32.0 GHz and 33.7 GHz) and at X-band (8.4 GHz), TMO Progress Report 42-125. Jet Propulsion Laboratory, Pasadena, CA, 1996.
57. Vlnrotter, V.A., and B. Iijima. Analysis of Array Feed Combining Performance Using Recorded Data, TMO Progress Report 42-125. Jet Propulsion Laboratory, Pasadena, CA, 1996.
58. Sue, M. Ka-band for Pluto fast flyby. *DSN Technol. Program News*, Issue 2, Jet Propulsion Laboratory, Pasadena, CA, 1995.

59. Gatti, M., and P. Stanton. Simultaneous X/Ka-band demonstration will benefit Cassini and others in the DSN. *DSN Technol. Sci. Program News*, Issue 7, Jet Propulsion Laboratory, Pasadena, CA, 1997.
60. Mysoor, N.R., et al. Performance of a Ka-band Transponder Breadboard for Deep Space Applications, TMO Progress Report 42-122. Jet Propulsion Laboratory, Pasadena, CA, 1995.
61. Lesh, J.R., J. Katz, H.H. Tan, and D. Zwillinger. 2.5 Bit/Detected Photon Demonstration Program: Description, Analysis, and Phase I Results, TDA Progress Report 42-66. Jet Propulsion Laboratory, Pasadena, CA, 1981, pp. 115–132.
62. Katz, J. 2.5 Bit/Detected Photon Demonstration Program: Phase II and III Experimental Results, TDA Progress Report 42-70. Jet Propulsion Laboratory, Pasadena, CA, 1982, pp. 95–104.
63. Wilson, K.E., J.R. Lesh, et al. GOPEX: A Deep Space Optical Communications Demonstration with the Galileo Spacecraft, TDA Progress Report 42-103. Jet Propulsion Laboratory, Pasadena, CA, 1990, pp. 262–273.
64. Wilson, K.E. An Overview of the GOLD Experiment Between the ETS VI Satellite and the Table Mountain Facility, TDA Progress Report 42-124. Jet Propulsion Laboratory, Pasadena, CA, 1996, pp. 8–19.
65. Ostro, S. Goldstone radar research of near-earth asteroids. *DSN Technol. Sci. Program News*, Issue 7, Jet Propulsion Laboratory, Pasadena, CA, 1997.

JAMES W. LAYLAND
LAWRENCE L. RAUCH
DOUGLAS J. MUDGWAY
JAMES R. LESH
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

E

EARTH ORBITING SATELLITE THEORY

Introduction

The objective of this article is to provide an introduction to the theory of satellite orbital motion. Although the discussion focuses on Earth-orbiting satellites, the theory is equally valid for computations around other primary bodies including the Moon, the Sun, and the planets. The majority of the discussion focuses on the problem of two bodies. The solution provides sufficient accuracy for many problems in astrodynamics. Perturbations from the two-body problem are discussed because a number of applications require greater accuracy. The solution of these more complex problems is generally accomplished using numerical integration techniques, which is beyond the scope of this chapter. The reader is encouraged to study References 1–7 for a more in-depth treatment of these problems.

Basic Principles of Orbital Motion

The basic principles governing the orbital motion of satellites were first formulated by Johannes Kepler and Isaac Newton in the 1600s. In 1619, Kepler, using the very precise observational data of Tycho Brahe, published what have become known as Kepler's laws:

1. The orbit of each planet is an ellipse where the Sun is at one focus.
2. The line joining the planet to the Sun sweeps out equal areas in equal times.
3. The square of the period of a planet is proportional to the cube of its mean distance to the Sun.

Kepler's laws described the kinematics of the motion of the planets, but the dynamic principles describing this motion were not solved until Isaac Newton published the *Principia* in 1687. Newton's three laws of motion are given as:

1. Every body continues in its state of rest or of uniform motion in a straight line, unless it is compelled to change that state by forces impressed upon it.
2. The change of motion is proportional to the motive force impressed and is made in the direction of the straight line in which that force is impressed.
3. To every action, there is always opposed an equal reaction, or the mutual actions of two bodies upon each other are always equal and directed to contrary parts.

The *Principia* also presents Newton's universal law of gravitation which can be stated as "any two point masses attract one another with a force proportional to the product of their masses and inversely proportional to the square of the distance between them." For a satellite orbiting the Earth, this law can be described mathematically as

$$\vec{F} = -\frac{GMm}{r^2} \left(\frac{\vec{r}}{r} \right), \quad (1)$$

where $G = 6.673 \times 10^{-20} \text{ km}^3/\text{kg s}^2$ is the universal gravitational constant, M is the mass of Earth, m is the mass of the satellite, \vec{r} is a vector from the center of mass of Earth to the center of mass of the satellite (both of which are assumed to be point masses here), and \vec{F} is the force acting on the satellite along \vec{r} .

These laws form the foundation upon which our understanding of satellite motion is built. Albert Einstein later explained small differences between observations and Newtonian dynamics, but Newton's laws provide a remarkably accurate model of satellite motion. The laws include some important concepts that will be used in the remainder of this chapter, including the concept of a *point mass* m , where all of the mass of a body is concentrated at a single point (or equivalently a uniform sphere), and the concept of an *inertial reference frame*. The concept of a force \vec{F} and variable time t are also fundamental to the laws of motion.

Two-Body Problem. Newton's laws of motion and law of gravitation allow us to derive the equations of motion of the satellite with respect to Earth. Using Newton's law of gravitation and his second law of motion, we can write the equations of motion of Earth and a satellite in an arbitrary inertial reference frame (Fig. 1) as

$$\vec{F}_{\text{sat}} = m\ddot{\vec{r}}_{\text{sat}} = -\frac{GMm}{r^2} \left(\frac{\vec{r}}{r} \right), \quad (2)$$

$$\vec{F}_{\text{Earth}} = M\ddot{\vec{r}}_{\text{Earth}} = \frac{GMm}{r^2} \left(\frac{\vec{r}}{r} \right). \quad (3)$$

We can then subtract one of these equations from the other to obtain the equations of motion of the satellite with respect to Earth (but still in the

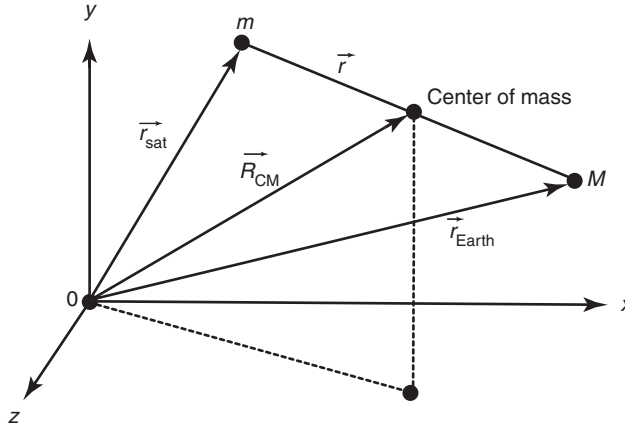


Figure 1. Derivation of the two-body problem (Fig. 3-1, Reference 6).

inertial frame):

$$\ddot{\vec{r}} = \ddot{\vec{r}}_{\text{sat}} - \ddot{\vec{r}}_{\text{Earth}} = -\frac{Gm}{r^2} \left(\frac{\vec{r}}{r} \right) - \frac{GM}{r^2} \left(\frac{\vec{r}}{r} \right) = -\frac{G(m+M)}{r^2} \left(\frac{\vec{r}}{r} \right). \quad (4)$$

It is common to ignore the mass of the satellite (as it is quite small with respect to the mass of Earth) and to define the gravitational constant $\mu = GM = 398600.44 \text{ km}^3/\text{s}^2$, so that Equation 4 may be written as

$$\ddot{\vec{r}} = -\frac{\mu \vec{r}}{r^3}, \quad (5)$$

which is a nonlinear, second-order, vector differential equation, often referred to as the “two-body equation.” It is important to remember the assumptions behind this equation:

1. The mass of the satellite is negligible with respect to the mass of the primary body (Earth).
2. The coordinate system is inertial.
3. Each of the two bodies behaves as a point mass (i.e., they are spherically symmetric and have uniform density).
4. Only the gravitational force between the two bodies is considered; no other forces (atmospheric drag, solar radiation pressure, or gravitational influence of other planets) are present.

Now, we will proceed to develop a solution to this equation, so that we may describe the shape of the satellite orbit about Earth and the position of the satellite in this orbit as a function of time. Let us examine the motion

of the center of mass of the two-body system. The definition of the center of mass is

$$\vec{R}_{\text{CM}} = \frac{M\vec{r}_{\text{Earth}} + m\vec{r}_{\text{sat}}}{M + m}. \quad (6)$$

Combining Equations 2,3 and 6 gives

$$\ddot{\vec{R}}_{\text{CM}} = 0 \rightarrow \ddot{\vec{R}}_{\text{CM}} = \vec{A}t + \vec{B}, \quad (7)$$

where \vec{A} and \vec{B} are constants of motion. Taking the vector cross-product of \vec{r} with Equation 5, it may be shown that an object subject only to two-body dynamics has a constant angular momentum vector:

$$\vec{r} \times \ddot{\vec{r}} = 0 \rightarrow \frac{d}{dt} (\vec{r} \times \dot{\vec{r}}) = 0 \rightarrow \vec{h} = \vec{r} \times \dot{\vec{r}} = \text{const.} \quad (8)$$

Thus, the angular momentum vector is constant, and the motion of m with respect to M is planar. Taking the dot product of $\dot{\vec{r}}$ with Equation 5, we can also show that the energy per unit mass of the motion of m with respect to M is a constant:

$$\xi = \frac{v^2}{2} - \frac{\mu}{r} = \text{const.}, \quad (9)$$

where r and v are the magnitudes of the position and velocity vectors, respectively. Also note that this equation represents the sum of the kinetic and potential energy. In summary, we have found 10 *integrals of motion* ($\vec{A}, \vec{B}, \vec{h}, \xi$), and we need to find two more to solve the two-body problem completely.

The two-body expression, Equation 5, is one of the few dynamic equations in orbital mechanics that can be solved analytically. Because the orbit of the satellite is planar, we can search for two-dimensional solutions to the problem. It is convenient to express Equation 5 in terms of polar coordinates, as shown in Fig. 2 (for an alternate derivation, see Reference 7). Using the following definitions,

$$\begin{aligned} \vec{r} &= r\vec{u}_r, \\ \dot{\vec{r}} &= \dot{r}\vec{u}_r + r\dot{\theta}\vec{u}_\theta, \\ \ddot{\vec{r}} &= \ddot{r}\vec{u}_r + \dot{r}\dot{\theta}\vec{u}_\theta + \dot{r}\ddot{\theta}\vec{u}_\theta + r\ddot{\theta}\vec{u}_\theta - r\dot{\theta}^2\vec{u}_r \\ &= (\ddot{r} - r\dot{\theta}^2)\vec{u}_r + (2\dot{r}\dot{\theta} + r\ddot{\theta})\vec{u}_\theta, \end{aligned} \quad (10)$$

we can rewrite Equation 5 as

$$r \text{ component: } \ddot{r} - r\dot{\theta}^2 = -\frac{\mu}{r^2}, \quad (11)$$

$$\theta \text{ component: } 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0. \quad (12)$$

First note that Equation 12 can be expressed as

$$\frac{d}{dt} (r^2\dot{\theta}) = 0. \quad (13)$$

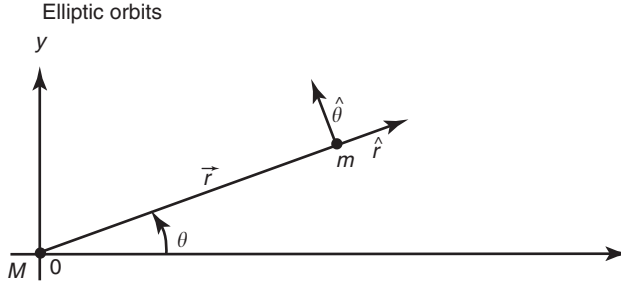


Figure 2. Use of polar coordinates for the two-body problem (Fig. 4-1, Reference 6).

Because $h = \vec{r} \times \dot{\vec{r}} = r^2 \dot{\theta} \vec{u}_h$ which is equal to a constant, from Equation 13, this again illustrates the conservation of angular momentum. The solution to Equation 11 is best found by making a change of variables of $r = 1/q$ and $t = \theta$ which gives the following differential equation:

$$\frac{d^2 q}{d\theta^2} + q = \frac{\mu}{h^2}. \quad (14)$$

This is a simple harmonic oscillator with a solution of

$$q = \frac{\mu}{h^2} + K \cos(\theta - \omega), \quad (15)$$

where K and ω are our last two integrals of motion. It can be shown that $K = \mu/eh^2$, where

$$e = \sqrt{1 + \frac{2\zeta h^2}{\mu^2}}. \quad (16)$$

Thus, after transforming variables again, Equation 15 can be written as

$$r = \frac{h^2/\mu}{1 + e \cos(\theta - \omega)}. \quad (17)$$

It can be shown that when $\theta - \omega = 0$, r is a minimum (called *periapse*), and when $\theta - \omega = \pi$, r is a maximum (*apoapse*). Thus, ω is referred to as the *argument of periapse*, and the location of the satellite relative to periapse is defined by the *true anomaly* $f = \theta - \omega$. Thus, using $p = h^2/\mu = a(1 - e^2)$ (called the *semiparameter* or *semilatus rectum*), the final solution to the two-body problem can be expressed as

$$r = \frac{p}{1 + e \cos f}, \quad (18)$$

which is the equation of a conic section, either an ellipse, a parabola, or a hyperbola. The size of the conic section is described by the semimajor axis a , and the shape of the conic section is described by its eccentricity e . The position of the satellite on the conic section is described by the true anomaly f , which is the angle of the satellite with respect to periapse (the point of closest approach to the planet). A satellite on an elliptical orbit ($0 < e < 1, a > 0$) has been captured by

the primary body and will not leave unless otherwise perturbed, and a satellite on a hyperbolic orbit ($e > 1, a < 0$), such as a planetary flyby, will never return to the primary body. A parabolic orbit ($e = 1, a = \infty$) is similar to a hyperbolic orbit and not normally encountered except in theory, and a circular orbit ($e = 0$) is a special case of an elliptical orbit.

For an elliptical orbit, there are some useful equations relating the semimajor axis a , the semiminor axis b , and the periapse (r_p) and apoapse (r_a) distances:

$$\begin{aligned} b &= a\sqrt{1 - e^2}, \\ r_p &= a(1 - e), \\ r_a &= a(1 + e). \end{aligned} \quad (19)$$

An alternative form can be derived for the energy of the orbit in terms of the semimajor axis:

$$\zeta = \frac{-\mu}{2a}, \quad (20)$$

which can be combined with Equation 9 to give a useful expression for the velocity:

$$V = \sqrt{\mu \left(\frac{2}{r} - \frac{1}{a} \right)}. \quad (21)$$

We can evaluate this equation for $r = a$ (circular orbit), $r = r_p$, and $r = r_a$, to get equations for the velocity of a circular orbit (V_c), the velocity at periapse (V_p), and the velocity at apoapse (V_a):

$$\begin{aligned} V_c &= \sqrt{\frac{\mu}{r}}, \\ V_p &= \sqrt{\frac{\mu(1 + e)}{a(1 - e)}}, \\ V_a &= \sqrt{\frac{\mu(1 - e)}{a(1 + e)}}. \end{aligned} \quad (22)$$

For an elliptical orbit, we can derive an expression for the period P of the orbit that is, the time it takes the satellite to make one revolution in its orbit. From Equation 13, we know that the magnitude of the angular momentum vector is

$$h = r^2 \frac{df}{dt}, \quad (23)$$

but we also know from calculus that the area swept out by the radius vector as it moves through a given angle is given by

$$dA = \frac{1}{2} r^2 df, \quad (24)$$

which, combined with the previous equation, gives

$$dt = \frac{2}{h} dA, \quad (25)$$

which proves Kepler's second law because h is constant. Integrating this equation over 2π radians of f gives an expression for the orbital period:

$$P = \frac{2\pi ab}{h}. \quad (26)$$

Then, if we substitute the relations $h = \sqrt{\mu p}$ and $b = \sqrt{ap}$ to obtain the most commonly used representation for the orbital period,

$$P = 2\pi \sqrt{\frac{a^3}{\mu}}, \quad (27)$$

which is also an expression of Kepler's third law.

Keplerian Orbit Elements. For an elliptical orbit, the semimajor axis a and the eccentricity e describe the size and shape of the ellipse, respectively, and the true anomaly determines the position of the satellite on the ellipse (Fig. 3). To determine the orientation of the ellipse in inertial space, we need to define three new angles, as shown in Fig. 4. The inclination i of the orbit is the angle between the angular momentum vector and the z axis of the coordinate system (K) (the

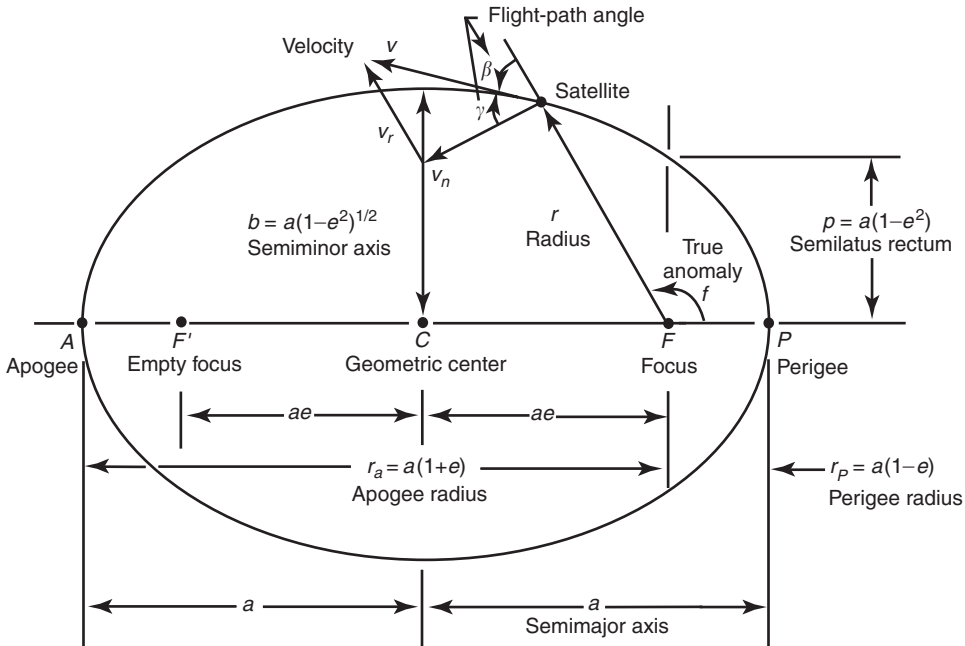


Figure 3. Diagram of an elliptical orbit (Fig. 3.3, Reference 2).

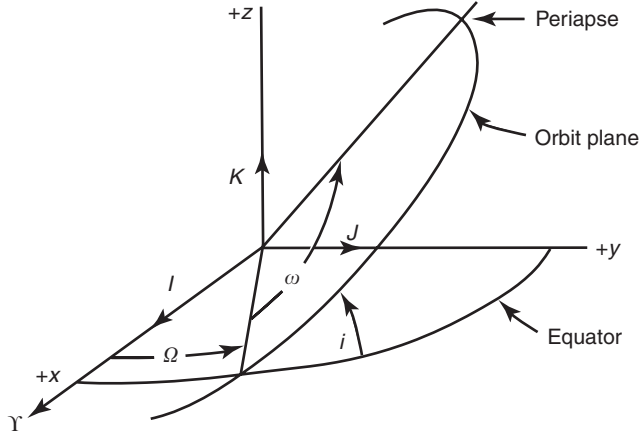


Figure 4. Definition of orbital elements (Fig. 3-6, Reference 15).

“tilt” of the orbital plane with respect to the equator). The longitude of the ascending node Ω is the angle between the line of nodes (the intersection of the plane of the orbit with the equatorial plane) and the x axis of the coordinate system. Finally, the argument of periape ω is the angle between the eccentricity vector (the vector that points toward periape) and the line of nodes. Together, a , e , i , Ω , ω and f at some time t constitute a set of Keplerian orbit elements that describe perfectly the motion of the satellite around a point mass planet (when coupled with Kepler’s equation in the next section). They can be considered equivalent to knowing the vector position and velocity at some time epoch, which can be used to determine the motion of the satellite by integrating Equation 5. Later, we will discuss how to determine orbit elements from position and velocity vectors and vice versa.

Although not an orbital element, another useful quantity to compute is the *flight-path angle*. The flight-path angle is the angle between the velocity vector and the local horizontal (as defined by a perpendicular to the position vector), and for noncircular orbits (circular orbits always have a flight-path angle of 0°), the flight-path angle is positive when the satellite is between periape and apoapse, and negative between apoapse and periape). It can be computed in the following manner:

$$\begin{aligned}\cos(\phi_{\text{fpa}}) &= \frac{1 + e \cos v}{\sqrt{1 + 2e \cos v + e^2}}, \\ \sin(\phi_{\text{fpa}}) &= \frac{e \sin v}{\sqrt{1 + 2e \cos v + e^2}}, \\ \phi_{\text{fpa}} &= \tan^{-1} \left(\frac{\sin \phi_{\text{fpa}}}{\cos \phi_{\text{fpa}}} \right),\end{aligned}\tag{28}$$

where the sine and cosine are explicitly given, so that the inverse tangent may be easily computed without a quadrant check.

Kepler's Equation. Kepler's equation allows us to determine the angular position of the satellite in its orbit as a function of time. We will not derive Kepler's equation here (see Reference 7 among others), however, this can be easily done using Kepler's second law and by defining a new position angle, the eccentric anomaly E , shown in Fig. 5. The eccentric anomaly is the angle between periapse and the satellite position projected onto an auxiliary circle with respect to the center of the ellipse (Fig. 5). Kepler's equation relates the eccentric anomaly to the mean anomaly M as

$$M = E - e \sin E \quad (29)$$

The mean anomaly is not a physical angle like the eccentric and true anomalies; it can only be defined mathematically:

$$M = n(t - t_p), \quad (30)$$

where t is time, t_p is the time of the satellite passage of periapse, and n is the mean motion, or the mean angular velocity of the satellite. The mean angular velocity of the satellite around one orbit is simply the orbital period divided by 2π :

$$n = \frac{P}{2\pi} = \sqrt{\frac{\mu}{a^3}}. \quad (31)$$

Thus, knowing the time past periapse passage and the semimajor axis of the orbit, we can compute the mean anomaly. Using the orbit eccentricity, we can then compute the eccentric anomaly from Kepler's equation. Finally, the true anomaly can be computed from the eccentric anomaly as

$$\tan\left(\frac{f}{2}\right) = \sqrt{\frac{1+e}{1-e}} \tan\left(\frac{E}{2}\right). \quad (32)$$

Note that this formulation is valid only for eccentric orbits; different formulations are needed for hyperbolic and parabolic orbits. Also, determining Kepler's equation for the eccentric anomaly is a nonlinear process, usually accomplished using

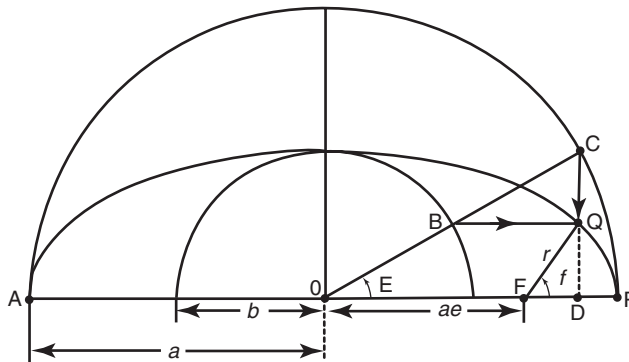


Figure 5. Definition of the eccentric anomaly (Fig. 7-1, Reference 6).

the Newton–Raphson iteration. If we define a function $f(E) = M - E + e \sin E = 0$, we can then expand this function in a Taylor series about some initial guess E_0 as

$$0 = f(E) = f(E_0 + \delta) = f(E_0) + f'(E_0)\delta + \text{higher order terms.} \quad (33)$$

Ignoring higher order terms, we can solve for δ as

$$\delta = \frac{f(E_0)}{f'(E_0)}. \quad (34)$$

Because we are ignoring higher order terms, we must iterate the solution for δ , and thus a general solution after taking the indicated derivative in Equation 34 is

$$E_{i+1} = E_i + \frac{M - E_i + e \sin(E_i)}{1 - e \cos(E_i)} \quad (35)$$

which should be iterated until $|E_{i+1} - E_i|$ is less than some prescribed tolerance.

Now, we have the tools to determine the satellite position as a function of time. Of the six Keplerian elements, only f varies with time (for two-body dynamics). At a given time t , we can compute M from Equation 30, and then E and f from Equations 29 and 32.

Computing Orbital Elements From Position and Velocity and Vice Versa. Computations in astrodynamics often require the ability to compute orbital elements from position and velocity vectors and vice versa. This arises because the satellite's orbital position as a function of time is most easily determined using orbital elements (in an analytical sense—numerical integration techniques work equally well with position and velocity), but other computations, such as determining satellite visibility at a ground station, are most easily accomplished using the position vector.

Until this point, we have not talked in detail about coordinate system definitions, but before we discuss the calculation of positional and velocity vectors, we must first define the inertial coordinate system in which they are represented. For Earth-orbiting applications, this is usually accomplished by using a geocentric equatorial system (Fig. 6), denoted here as Earth-centered inertial (ECI) or IJK . As shown in Fig. 6, the fundamental plane (IJ) is Earth's equatorial plane. The I axis points towards the vernal equinox on a particular date (the position of the Sun against the stars as it ascends across the equator near March 21); the K axis points toward the North Pole, and the J axis completes the right-handed system, 90° to the I axis. Currently, the J2000 system, which is a geocentric equatorial frame defined at the J2000 epoch (noon on 1 January 2000), is the most commonly used ECI reference frame. The ECI frame is by definition non-rotating, because it is quasi-inertial. However, for many applications, such as describing the location of tracking stations, we also need to define a body-fixed rotating reference frame. This is typically referred to as the Earth-centered Earth-fixed (ECEF) frame (IJK_{ECEF}), where the I_{ECEF} axis is pointed at 0° longitude on the equator (the Greenwich meridian), the K_{ECEF} axis points at the North Pole, and the J_{ECEF} axis completes the right-handed system (points at 90° longitude). In the absence of the relatively small effects of precession, nutation, and polar motion, the K and K_{ECEF} axes are equivalent, and the IJ and IJ_{ECEF}

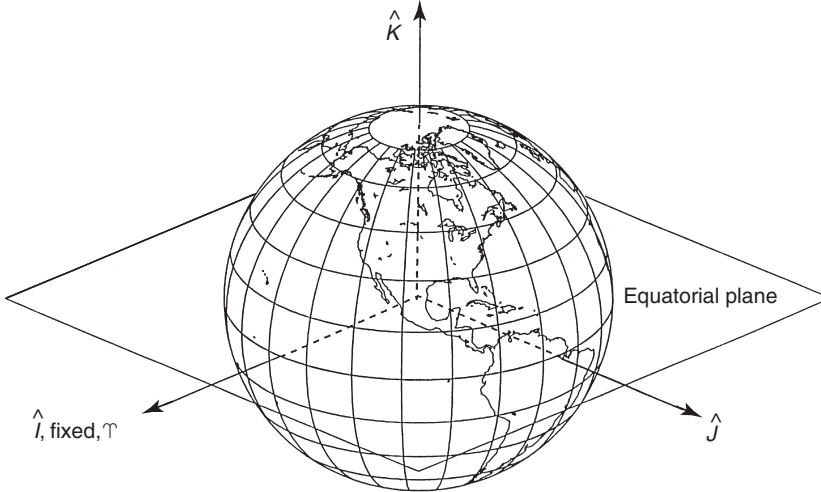


Figure 6. Coordinates systems used in the orbit problem (Fig. 1-20, Reference 7).

axes are related through Greenwich sidereal time θ_{GST} (Fig. 7), which is the angle between the I axis and the Greenwich meridian (0° longitude). This angle varies between 0° and 360° as the Earth rotates. Additional rotations are required to account for the effects of precession, nutation, and polar motion (see Reference 7).

Converting between the ECI and ECEF frames, as well as other conversions we will discuss later, can be accomplished using principal axis rotations which are defined as follows:

$$\begin{aligned}
 \text{ROTX}(\theta) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{bmatrix} \\
 \text{ROTY}(\theta) &= \begin{bmatrix} \cos \theta & 0 & -\sin \theta \\ 0 & 1 & 0 \\ \sin \theta & 0 & \cos \theta \end{bmatrix} \\
 \text{ROTZ}(\theta) &= \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}
 \end{aligned} \tag{36}$$

where each matrix, when used to premultiply a given vector, will rotate that vector by an angle θ around the indicated axis. Thus, ignoring precession, nutation, and polar motion, the transformations from ECI to ECEF and back are as follows:

$$\begin{aligned}
 \vec{r}_{\text{ECEF}} &= \text{ROTZ}(\theta_{\text{GST}}) \vec{r}_{\text{ECI}}, \\
 \vec{r}_{\text{ECI}} &= \text{ROTZ}(-\theta_{\text{GST}}) \vec{r}_{\text{ECEF}}.
 \end{aligned} \tag{37}$$

The position of the vernal equinox changes in time, and thus a specific date must be assigned to the equinox used to define the reference frame. A commonly

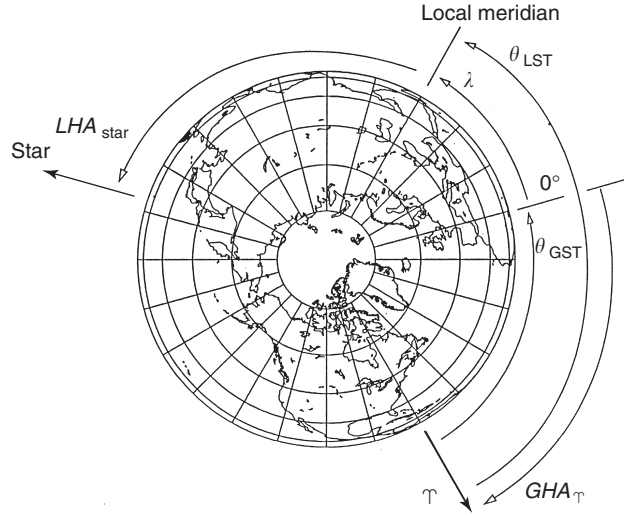


Figure 7. Relationship between the ECI and ECEF coordinate systems (Fig. 1-31, Reference 7).

employed reference is the mean equator and equinox of J2000.0. Mean sidereal time accounts for the precession of the equinox, and apparent sidereal time includes secular and periodic motions of the equinox; the difference is referred to as the equation of the equinoxes. Generally, we use Greenwich mean sidereal time, which can be approximately represented as

$$\begin{aligned}\theta_{\text{GST}} &= \theta_{\text{GST}0} + \omega_e \Delta t \\ \theta_{\text{GST}0} &= 1.753368560 + 628.3319706889 \Delta T \\ &\quad + 6.7707 \times 10^{-6} \Delta T^2 - 4.5 \times 10^{-10} \Delta T^3,\end{aligned}\tag{38}$$

where θ_{GST} is given in radians, ω_e is the rotational rate of Earth ($7.29211585530 \times 10^{-5}$ rad/s), ΔT is the number of Julian centuries elapsed from the epoch J2000.0 (1 January 2000, 12 hrs) for the day of interest, and Δt is the number of seconds from the beginning of the day.

Now, we are ready to describe the transformation of position/velocity to orbit elements and back. Given a vector position \vec{r} and velocity \vec{v} defined in the ECI frame, we can compute orbital elements using the following procedure.

Position/Velocity to Orbital Elements. The vectors used in this discussion are illustrated in Fig. 4. We can find the semimajor axis using the energy integral (Eq. 9) as follows:

$$a = \left(\frac{2}{r} - \frac{v^2}{\mu} \right)^{-1}.\tag{39}$$

For several reasons, it is convenient to define the eccentricity vector (which points at periapse, Fig. 4) as

$$\vec{e} = \frac{\vec{v} \times \vec{h}}{\mu} - \frac{\vec{r}}{r}.\tag{40}$$

The eccentricity of the orbit can be found from the magnitude of the eccentricity vector, or using Equation 16,

$$e = |\vec{e}| = \sqrt{1 + \frac{2\zeta h^2}{\mu^2}}. \quad (41)$$

The inclination i of the orbit, is the angle between the angular momentum vector and the K axis. Thus using the law of cosines ($ab \cos \alpha = \vec{a} \cdot \vec{b}$),

$$i = \cos^{-1} \left(\frac{\hat{K} \cdot \vec{h}}{|\hat{K}| |\vec{h}|} \right). \quad (42)$$

The longitude of the ascending node is the angle measured in the equatorial plane from the I axis to the satellite (i.e., it is the angle between I and the node vector $\vec{n} = \hat{K} \times \vec{h}$). Thus,

$$\Omega = \cos^{-1} \left(\frac{\hat{I} \cdot \vec{n}}{|\hat{I}| |\vec{n}|} \right) \quad (43)$$

The ascending node may be computed from Equation 43 if a quadrant check is performed (if the y component of the node vector is less than zero, the ascending node lies between 180 and 360°). The argument of periapse ω is the angle between the node vector and the eccentricity vector (Fig. 4). Thus,

$$\omega = \cos^{-1} \left(\frac{\vec{n} \cdot \vec{e}}{|\vec{n}| |\vec{e}|} \right), \quad (44)$$

where again the quadrant must be checked (if the z component of the eccentricity vector is negative, ω lies between 180 and 360°). Finally, the true anomaly is the angle between the position vector and the eccentricity vector.

$$f = \cos^{-1} \left(\frac{\vec{r} \cdot \vec{e}}{|\vec{r}| |\vec{e}|} \right), \quad (45)$$

where the quadrant must be checked using the flight-path angle (if $\phi_{\text{fpa}} < 0$, $180^\circ \leq \nu \leq 360^\circ$). Although these formulas are reasonably general, several special cases can arise that require a different approach, namely, circular and/or equatorial orbits (see Reference 7).

Orbital Elements to Position/Velocity. Orbital elements are best transformed to position/velocity by using the principal axis rotations defined earlier. First, we write the position and velocity vectors as expressed in a “perifocal” reference frame, which has x pointed at periapse, z normal to the orbit plane, and y completing the right-hand system (Fig. 8). In this frame, the position and

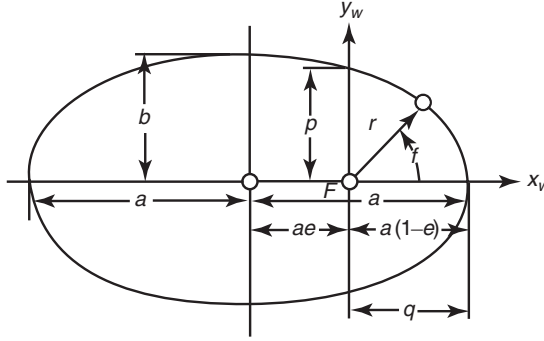


Figure 8. The perifocal reference frame (Fig. 3-3, Reference 15).

velocity can be written as a function of orbital elements:

$$\begin{aligned} \vec{r}_{\text{PER}} &= \begin{bmatrix} r \cos f \\ r \sin f \\ 0 \end{bmatrix}, \\ \vec{v}_{\text{PER}} &= \begin{bmatrix} -\sqrt{\frac{\mu}{p}} \sin f \\ \sqrt{\frac{\mu}{p}}(e + \cos f) \\ 0 \end{bmatrix}, \end{aligned} \quad (46)$$

where r is computed from Equation 18 and $p = a(1 - e^2)$ is the semiparameter. Once the position and velocity vectors are defined in the perifocal frame, they may be rotated to the ECI frame through simple rotations of $-\omega$ about z , $-i$ about x , and $-\Omega$ about z , as follows:

$$\begin{aligned} \vec{r}_{IJK} &= \text{ROTZ}(-\Omega)\text{ROTX}(-i)\text{ROTZ}(-\omega)\vec{r}_{\text{PER}}, \\ \vec{v}_{IJK} &= \text{ROTZ}(-\Omega)\text{ROTX}(-i)\text{ROTZ}(-\omega)\vec{v}_{\text{PER}}. \end{aligned} \quad (47)$$

In summary, the equations of motion for the two-body problem can be solved and yield a conic section that is best described using Keplerian orbital elements. Given a Cartesian position and velocity for a satellite at a given time, a set of orbital elements may be computed using the preceding equations. These elements may be easily propagated in time because only M varies in time via Kepler's equation. The propagated elements may be converted back to inertial Cartesian position and velocity using Equations 46 and 47 and into the ECEF frame via Equations 37. The Cartesian position in the ECEF frame is related to geocentric latitude ϕ , longitude λ and radius r as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} r \cos \phi \cos \lambda \\ r \cos \phi \sin \lambda \\ r \sin \phi \end{bmatrix}. \quad (48)$$

Thus, the latitude and longitude of the satellite may be computed as

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2}, \\ \phi &= \sin^{-1} \frac{z}{r}, \\ \lambda &= \tan^{-1} \frac{y}{x}. \end{aligned} \quad (49)$$

It is sometimes also desirable to compute the position of the satellite relative to some fixed point on Earth, such as a tracking site. This can be done by defining a site-fixed topocentric south-east-up (SEZ) reference frame, where the x axis is pointed south, the y axis is pointed east, and the z axis points toward the zenith. Given the Cartesian position of the satellite relative to the site in ECEF frame,

$$\begin{aligned} \vec{\rho}_{\text{ECEF}} &= \vec{r}_{\text{ECEF}} - \vec{r}_{\text{siteECEF}}, \\ \vec{r}_{\text{siteECEF}} &= \begin{bmatrix} r_{\text{site}} \cos \phi_{\text{site}} \cos \lambda_{\text{site}} \\ r_{\text{site}} \cos \phi_{\text{site}} \sin \lambda_{\text{site}} \\ r_{\text{site}} \sin \phi_{\text{site}} \end{bmatrix}. \end{aligned} \quad (50)$$

The transformation of a vector from the ECEF frame to the SEZ frame is given as

$$\vec{\rho}_{\text{SEZ}} = \text{ROTY}(90^\circ - \phi_{\text{site}}) \text{ROTZ}(\lambda_{\text{site}}) \vec{\rho}_{\text{ECEF}}. \quad (51)$$

The azimuth and elevation of the satellite relative to the site are related to the position in the topocentric frame via

$$\vec{\rho}_{\text{SEZ}} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -\rho \cos el \cos az \\ \rho \cos el \sin az \\ \rho \sin el \end{bmatrix}, \quad (52)$$

and thus the azimuth and elevation may be easily computed as

$$\begin{aligned} \rho &= \sqrt{x^2 + y^2 + z^2}, \\ el &= \sin^{-1} \frac{z}{\rho}, \\ az &= \tan^{-1} \frac{y}{-x}. \end{aligned} \quad (53)$$

Perturbation Theory

The discussion to this point has assumed that Earth can be represented by a point mass, that is, a spherically symmetrical homogeneous body, and that no nongravitational forces act on the satellite. In reality, Earth's shape and mass deviate significantly from a sphere, and a satellite is subject to the significant perturbations of atmospheric drag and solar radiative pressure among other nongravitational forces. The influence of these other effects on the orbit will now be described.

Perturbations Due to Earth's Nonspherical Shape. The shape of Earth and the distribution of mass within it are quite complex. Normally, we describe Earth's gravity field in terms of its potential, and then take the gradient of the potential to determine the force on the satellite. For example, the two-body equations, Equation 5, can be expressed in this way:

$$\begin{aligned}\ddot{\mathbf{r}} &= \nabla U, \\ U &= \frac{\mu}{r},\end{aligned}\tag{54}$$

where ∇U is the gradient of the potential U . The potential shown here includes only a point mass term. Deviations of the gravity field from a point mass are described by using an expansion of the potential in terms of spherical harmonic functions as

$$\begin{aligned}U &= \frac{\mu}{r} \left[1 - \sum_{l=2}^{\infty} J_l \left(\frac{R_{\oplus}}{r} \right)^l P_l(\sin \phi_{\text{sat}}) \right. \\ &\quad + \sum_{l=2}^{\infty} \sum_{m=l}^l \left(\frac{R_{\oplus}}{r} \right)^l P_{lm}(\sin \phi_{\text{sat}}) (C_{lm} \cos m\lambda_{\text{sat}} \\ &\quad \left. + S_{lm} \sin m\lambda_{\text{sat}}) \right],\end{aligned}\tag{55}$$

where P_l are the Legendre polynomials of degree l , P_{lm} are the Legendre associated functions of degree l and order m , r is the magnitude of the satellite position vector, ϕ_{sat} and λ_{sat} are the satellite latitude and longitude, R_{\oplus} is the radius of Earth (~ 6378 km), J_l are zonal gravitational coefficients describing the latitudinal gravity variations, and C_{lm}/S_{lm} are the sectorial ($l=m$) and tesseral ($l \neq m \neq 0$) gravitational coefficients describing the longitudinal gravity variations. The gravitational coefficients are a function of the mass distribution within Earth. The $l=2$ terms are related to the moments and products of inertia of Earth as

$$\begin{aligned}J_2 &= \frac{2I_{zz} - I_{yy} - I_{xx}}{2MR_{\oplus}^2}, \\ C_{2,1} &= \frac{I_{yz}}{MR_{\oplus}^2}, \quad S_{2,1} = \frac{I_{xz}}{MR_{\oplus}^2}, \\ C_{2,2} &= \frac{I_{yy} - I_{xx}}{4MR_{\oplus}^2}, \quad S_{2,2} = \frac{I_{xy}}{2MR_{\oplus}^2}.\end{aligned}\tag{56}$$

The J_2 term describes the oblateness of Earth, and the J_3 coefficient describes its slight “pear shape.” The $l=1$ terms are normally considered zero if the origin of the coordinate system coincides with Earth's center of mass.

Note that taking the gradient of U in Equation 54 is not trivial because U is expressed in terms of a body-fixed position of the satellite r , ϕ , and λ . Thus, the

gradient should be computed in spherical coordinates as

$$\nabla U = \frac{\partial U}{\partial r} \vec{u}_r + \frac{1}{r} \frac{\partial U}{\partial \phi} \vec{u}_\phi + \frac{1}{r \cos \phi} \frac{\partial U}{\partial \lambda} \vec{u}_\lambda. \quad (57)$$

Then, we must rotate this vector quantity from the satellite-fixed polar coordinates to ECEF coordinates and finally to JK inertial coordinates. Thus the final expression for the force of the gravity field can be expressed as

$$\vec{F}_g = \text{ROT3}(-\theta_{\text{gst}}) \text{ROTZ}(-\lambda) \text{ROTY}(\phi) \nabla U. \quad (58)$$

Note that the first term in our expression of the potential (Equation 55) is just the point mass potential, and we know how the orbit behaves in the presence of a point mass. Thus, it is customary to define the *disturbing potential* and the resulting force due to the nonsphericity of the gravity field as

$$R = U - \frac{\mu}{r}, \quad (59)$$

$$\vec{F}_{\text{ns}} = \text{ROTZ}(-\theta_{\text{gst}}) \text{ROTZ}(-\lambda) \text{ROTY}(\phi) \nabla R.$$

The gravitational coefficients of Earth (J_l , C_{lm} , S_{lm}) have been determined from tracking observations of artificial Earth satellites (see Reference 8 for a review). The orbital perturbations due to Earth's gravity field can be classified as either secular (varying linearly in time), short period (approximately the satellite orbital period), and long period (weeks to months). All of these perturbations can be important for some applications, but we will restrict our discussion to secular perturbations because they do not average zero over time. The largest coefficient in the spherical harmonic expansion is the J_2 term, which describes the ellipsoidal shape of Earth. Now, we need to describe the technique for determining the effect of J_2 on the orbit.

General versus Special Perturbations. To determine how a given perturbative force acts on a satellite, one could perform a straight numerical integration of the equations of motion on a computer, referred to in the astrodynamics community as *special perturbations*, or one could analyze the perturbations using analytical techniques, referred to as *general perturbations*. The former technique is quite useful for propagating orbits with complex dynamics that do not lend themselves to an analytical treatment, or when an exact representation is required. The latter technique often leads to a more thorough understanding of the effects on the orbit, at the expense of inaccuracies due to linearization of the forcing functions. Some forces are so complex that applying the general perturbation technique is quite unwieldy. The use of the Lagrange planetary equations is a particularly useful general perturbation technique for examining gravitational perturbations.

Lagrange Planetary Equations. Most general perturbation techniques rely on a method called *variation of parameters*. Euler and Lagrange initially developed this method to describe conservative (gravitational) accelerations, and Gauss' method also applies to nonconservative accelerations. Essentially, these techniques use the two-body orbit as a solution to an unperturbed system and

then linearize the perturbations around the unperturbed system. Equations can be developed that describe the secular change of the orbital elements from an unperturbed system. The Lagrange planetary equations present these equations as a function of the disturbing potential R as

$$\begin{aligned}
 \frac{da}{dt} &= \frac{2}{na} \frac{\partial R}{\partial M_0}, \\
 \frac{de}{dt} &= \frac{1-e^2}{na^2e} \frac{\partial R}{\partial M_0} - \frac{\sqrt{1-e^2}}{na^2e} \frac{\partial R}{\partial \omega}, \\
 \frac{di}{dt} &= \frac{1}{na^2\sqrt{1-e^2}\sin i} \left(\cos i \frac{\partial R}{\partial \omega} - \frac{\partial R}{\partial \Omega} \right), \\
 \frac{d\omega}{dt} &= \frac{\sqrt{1-e^2}}{na^2e} \frac{\partial R}{\partial e} - \frac{\cot i}{na^2\sqrt{1-e^2}} \frac{\partial R}{\partial i}, \\
 \frac{d\Omega}{dt} &= \frac{1}{na^2\sqrt{1-e^2}\sin i} \frac{\partial R}{\partial i}, \\
 \frac{dM_0}{dt} &= \frac{1-e^2}{na^2e} \frac{\partial R}{\partial e} - \frac{2}{na} \frac{\partial R}{\partial a}.
 \end{aligned} \tag{60}$$

Note that these equations require the partial derivatives of the disturbing potential with respect to the orbital elements, whereas the previous expression for the disturbing potential is in terms of satellite position (latitude, longitude, and height). Expressing the disturbing potential R in terms of orbital elements is described by Kaula (9), who used this with the Lagrange planetary equations to investigate gravitational perturbations to the orbit. As the manipulation of the equations is somewhat complex, we will examine the procedure only for J_2 .

J_2 Perturbations and the Secular Precessing Ellipse. As previously mentioned, the J_2 gravitational coefficient is by far the most significant term in the definition of the nonsphericity of Earth's gravity field. The secular effect of J_2 on the orbital elements can be determined by isolating this term in the disturbing potential and substituting it in the Lagrange planetary equations. The J_2 term of the disturbing potential is

$$R_{J_2} = -\frac{\mu J_2}{r} \left(\frac{R_\oplus}{r} \right)^2 \frac{3}{2} \left(\sin^2 \phi_{\text{sat}} - \frac{1}{3} \right). \tag{61}$$

This equation can be expressed in terms of orbital elements using $\sin \phi_{\text{sat}} = \sin i \sin(\omega + f)$ and a trigonometric identity. Because we are interested in secular effects, we can ignore terms containing ω and f . If we further average the disturbing potential over one orbital revolution, the final result is (7)

$$R_{J_2} = -\frac{3}{2} n^2 R_\oplus^2 J_2 \frac{1}{(1-e^2)^{\frac{3}{2}}} \left(\frac{\sin^2 i}{2} - \frac{1}{3} \right). \tag{62}$$

Then, this can be substituted in the Lagrange planetary equations (60) to yield (ignoring higher order terms in J_2)

$$\begin{aligned}
 \frac{da}{dt} &= 0, \\
 \frac{de}{dt} &= 0, \\
 \frac{di}{dt} &= 0, \\
 \frac{d\Omega}{dt} &= -\frac{3J_2 R_\oplus^2 n \cos i}{2p^2}, \\
 \frac{d\omega}{dt} &= \frac{3nR_\oplus^2 J_2}{4p^2} (4 - 5 \sin^2 i), \\
 \frac{dM_0}{dt} &= \frac{3nR_\oplus^2 J_2 \sqrt{1-e^2}}{4p^2} (3 \sin^2 i - 2).
 \end{aligned} \tag{63}$$

The last equation describes secular changes in the rate of change of the mean anomaly (i.e., changes in the mean motion). The secular changes in the longitude of the ascending node and the argument of periapse have led to the description

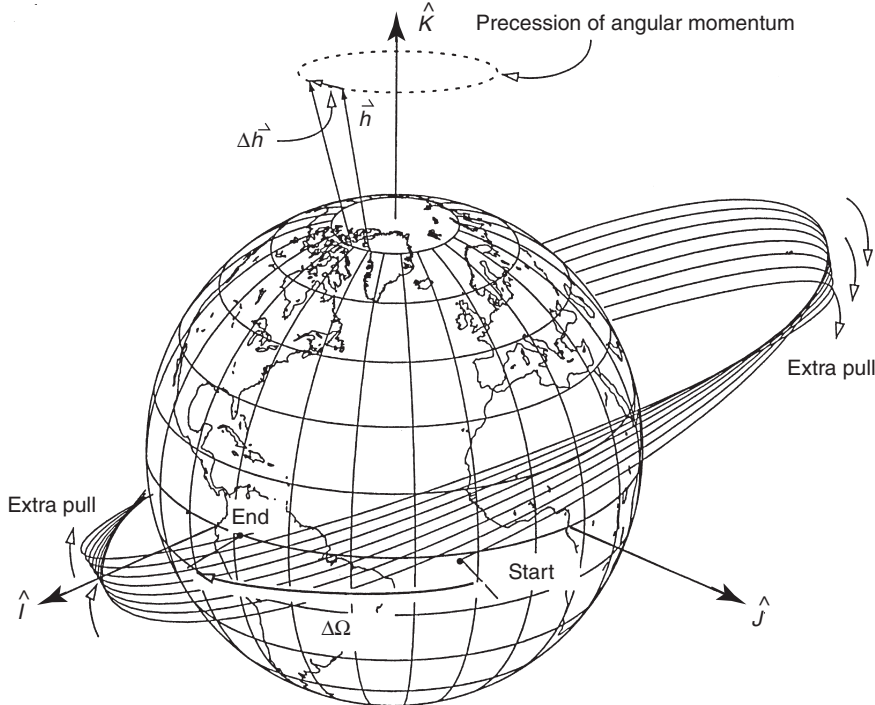


Figure 9. Precession of the longitude of the ascending node (Fig. 8-3, Reference 7).

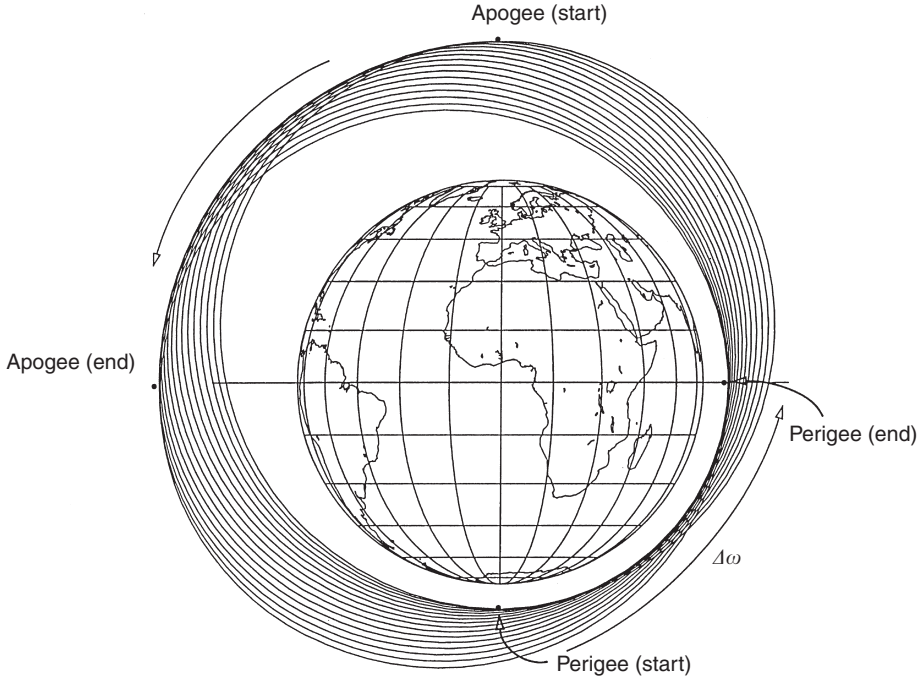


Figure 10. Precession of the argument of periapee (Fig. 8-5, Reference 7).

secular precessing ellipse, as the orbit precesses around the z axis (Fig. 9), and periapee moves around the orbit (Fig. 10). The rate of change of the node can vary by $\pm 10^\circ/\text{day}$, depending on the orbit characteristics. It is zero for a satellite at 90° inclination, and negative/positive for inclinations less/greater than 90° . The rate of change of the argument of periapee can have a similar magnitude and is zero for an inclination of 63.4° or 116.6° , called the *critical inclination*. Thus, under only the influence of J_2 , the secular change of the orbital elements can be written as

$$\begin{aligned}\Omega(t) &= \Omega(t_0) + \frac{d\Omega}{dt}(t - t_0), \\ \omega(t) &= \omega(t_0) + \frac{d\omega}{dt}(t - t_0), \\ M(t) &= M(t_0) + \left(\frac{dM_0}{dt} + n \right)(t - t_0),\end{aligned}\tag{64}$$

where the other orbital elements (a , e , i) remain constant.

The Lagrange planetary equations are useful only for evaluating conservative forces that can be represented in terms of a potential. There is a Gaussian form of the equations that can be used to study nonconservative forces (7). The evaluation of other perturbations using a general perturbation technique can become quite complex and is beyond the scope of this article. Thus, the remaining perturbations will be described in terms of their contribution to the two-body

equations of motion (5), under the assumption that a special perturbation technique (numerical integration) will be used to determine their effect on the orbit.

Perturbations Due to Drag and Solar Radiative Pressure. Atmospheric drag can be one of the largest perturbing forces for low altitude Earth satellites. The drag force per unit mass can be represented as

$$\vec{F}_{\text{drag}} = -\frac{1}{2}\rho\left(\frac{C_D A}{m}\right)V_r \vec{V}_r \quad (65)$$

where ρ is the atmospheric density, C_D is the drag coefficient, \vec{V}_r is the satellite velocity relative to the atmosphere, m is the mass of the satellite, and A is the cross-sectional satellite area. The quantity in brackets is sometimes referred to as the *ballistic coefficient*. The atmospheric density drops off approximately exponentially with increasing altitude, and it can have considerable temporal and spatial variation, due to the location of the subsolar point and solar and geomagnetic activity. Commonly used models for the atmospheric density include those in References 10–12. These models commonly use geomagnetic indexes and solar flux values to model the temporal variations of the density. The models are far from perfect, and deficiencies in the models are usually accounted for by estimating the drag coefficient in a piecewise continuous fashion using tracking data. Atmospheric drag is the leading factor that determines the lifetime of the satellite.

Atmospheric drag decreases the energy of the satellite over time, as manifested by decays in the semimajor axis and eccentricity of the orbit. For an eccentric orbit, the periape radius will remain roughly constant, whereas the apoapse radius decreases over time, corresponding to a decrease in the semimajor axis and the eccentricity of the orbit.

Radiative pressure is also an important force for precise modeling of satellite dynamics. Direct solar radiation is the largest effect, but reflected and emitted radiation from Earth can also be important. The force per unit mass due to solar radiative pressure is given by

$$\vec{F}_{\text{SRP}} = -P \frac{vA}{m} C_R \vec{u}_s, \quad (66)$$

where P is the momentum flux from the Sun, \vec{u}_s is the unit vector from the Sun to the satellite, A is the cross-sectional area in the direction of \vec{u}_s , C_R is the reflectivity coefficient whose values are between 1 and 2, and v is a shadow factor equal to 1 if the satellite is in sunlight, 0 if the satellite is in shadow (umbra), and between 0 and 1 if it is in partial shadow (penumbra). Earth radiative pressure can be modeled similarly; however, P would be modified to account for the spatial variations of albedo and emissivity over Earth's surface. In a manner analogous to drag, satellite reflectivity is sometimes estimated (using satellite tracking data) to account for errors in modeling solar radiative pressure.

Errors in the models for atmospheric drag and solar radiative pressure are quite complex, even when the drag and reflectivity coefficients are estimated. One commonly used method for representing these errors, as well as errors due to other nonconservative forces, is to estimate empirical forces acting on the

satellite. Because these orbit errors typically have a frequency of once per orbital period, the form of the empirical force is taken as $A + B \cos \alpha t + C \sin \alpha t$, where α is a frequency of once per revolution and the coefficients A , B , and C are estimated from the tracking data. This force is often represented by vector components in the radial, transverse, and normal directions.

N-Body Perturbations. It is often necessary to include the perturbative effects of other planetary bodies, such as the Moon, Jupiter, and Saturn, in the equations of motion for an Earth-orbiting satellite. This is usually done by considering each of the bodies as a point mass. Using an inertial coordinate system, as shown in Fig. 11, the equations of motion of the satellite with respect to Earth in the presence of a third body can be written as

$$\ddot{\vec{r}} = \ddot{\vec{r}}_{\text{sat}} - \ddot{\vec{r}}_{\oplus}. \quad (67)$$

Newton's second law and the law of gravitation give

$$\begin{aligned} M\ddot{\vec{r}}_{\oplus} &= \frac{GMm\vec{r}_{\oplus\text{sat}}}{r_{\oplus\text{sat}}^3} + \frac{GMm_{\odot}\vec{r}_{\oplus\odot}}{r_{\oplus\odot}^3}, \\ m\ddot{\vec{r}}_{\text{sat}} &= -\frac{GMm\vec{r}_{\oplus\text{sat}}}{r_{\oplus\text{sat}}^3} - \frac{Gm_{\odot}m\vec{r}_{\odot\text{sat}}}{r_{\odot\text{sat}}^3}, \end{aligned} \quad (68)$$

that can be substituted in Equation 67 to yield (after rearranging terms)

$$\ddot{\vec{r}} = -\frac{G(M+m)\vec{r}}{r^3} + Gm_{\odot} \left(\frac{\vec{r}_{\text{sat}\odot}}{r_{\text{sat}\odot}^3} - \frac{\vec{r}_{\oplus\odot}}{r_{\oplus\odot}^3} \right). \quad (69)$$

We recognize the first term in this equation as the two-body term from Equation 5. The second term represents the perturbation from two-body motion. It is composed of a *direct effect*, because it is the direct acceleration of the third body on the satellite, and an *indirect effect*, because it represents the perturbation of the third body on Earth and is not function of the satellite position. We can generalize this

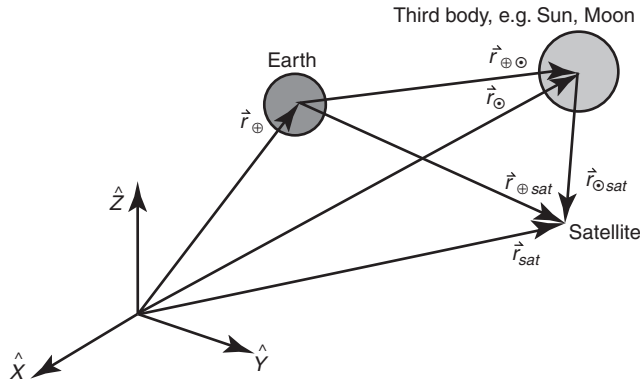


Figure 11. Perturbations due to a third body (Fig. 2-6, Reference 7).

equation from a single third body to N “third” bodies as

$$\begin{aligned}\ddot{\vec{r}}_{\oplus\text{sat}} &= -\frac{G(m_{\oplus} + m_{\text{sat}})\vec{r}_{\oplus\text{sat}}}{r_{\oplus\text{sat}}^3} + \vec{F}_{N\text{-body}} \\ \vec{F}_{N\text{-body}} &= G \sum_{j=1}^N m_j \left(\frac{\vec{r}_{\text{sat}j}}{r_{\text{sat}j}^3} - \frac{\vec{r}_{\oplus j}}{r_{\oplus j}^3} \right)\end{aligned}\quad (70)$$

This equation may be solved only by using numerical integration techniques (special perturbations), although several integrals do exist (7). For a single third body, the *restricted three-body problem* (13), where it is assumed that the satellite has no effect on the motion of the other two primary masses and the primary masses move on circular orbits, does have solutions.

Summary. For nonspherical gravitational perturbations only, the time rate of change of the orbital elements may be computed analytically using the Lagrange planetary equations, such as for the secular J_2 perturbations to Ω , ω , and M . However, if a precise representation is needed that includes the effect of non-gravitational forces, direct numerical integration of the equations of motion is best. The expanded equations of motion would be

$$\ddot{\vec{r}} = -\frac{\mu\vec{r}}{r^3} + \vec{F}_{\text{NS}} + \vec{F}_{\text{drag}} + \vec{F}_{\text{SRP}} + \vec{F}_{\text{ERP}} + \vec{F}_{N\text{-body}} + \vec{F}_{\text{other}}, \quad (71)$$

where \vec{F}_{other} might represent relativistic effects, satellite thermal venting, or magnetic effects. Note that for each force component, care must be taken to represent each in a common coordinate system, usually the inertial J2000.0 system.

Types of Orbits

Polar, Prograde, and Retrograde Orbits. Choosing the inclination of an orbit is a critical component of the mission design. Specifying the inclination of the orbit immediately determines the maximum latitude on Earth over which the satellite will pass. When considered in context with Earth’s rotation, the satellite ground track can evolve quite differently if the orbit is *prograde* versus *retrograde*. A prograde orbit has an inclination less than 90° (thus the satellite’s motion is in the same direction as Earth’s rotation), and its nodal precession rate ($\dot{\Omega}$) is negative. A retrograde orbit has an inclination greater than 90° (thus it moves in a direction opposite to Earth’s rotation), and its nodal precession rate is positive. A *polar* orbit has an inclination of 90° and is noteworthy because the orbit plane does not precess due to zonal gravitational perturbations ($\dot{\Omega} = 0$).

Equatorial and Geosynchronous Orbits. If the inclination of the orbit is 0° , then it is referred to as an *equatorial* orbit, and the satellite’s ground track never leaves the equator. A *geosynchronous* orbit has a period equal to Earth’s rotational period, which implies that it can be obtained by properly selecting the semimajor axis:

$$P = 2\pi\sqrt{\frac{a^3}{\mu}} = 1 \text{ sidereal day} \quad (72)$$

Solving for the semimajor axis gives $a \cong 42,164$ km to attain a geosynchronous orbit. If the satellite is also in an equatorial orbit, then it stays over the same geographical point on Earth at all times, and the orbit is referred to as *geostationary*. This is the orbit used by many communications satellites.

Sun-Synchronous Orbits. As the name implies, *Sun-synchronous* orbits maintain a constant orientation of the orbit plane with respect to the Sun throughout the year. This is accomplished by setting the orbit's nodal precession rate (caused by the perturbations of J_2) equal to the angular rate of Earth's orbit around the Sun of about a degree per day ($360^\circ/365.25$ days = 0.0172028 rad/s). Setting this equal to our equation for the nodal precession rate,

$$0.0172028 = -\frac{3nR_\oplus^2 J_2}{2p^2} \cos i, \quad (73)$$

gives an equation for determining the orbital parameters of a Sun-synchronous orbit. For example, for a circular orbit ($e=0.0$) at 800 km altitude ($a=p=7178$ km), solving for the inclination gives $i=98.6^\circ$. This is a common orbit for Earth observation satellites because the solar illumination conditions (i.e., local time) at Earth's surface are identical each time the satellite passes over the same latitude on the ground.

Other Orbits. A *frozen orbit* refers to a case where the secular variations of eccentricity and argument of perigee are designed to be zero, so that global variations in altitude are minimized. Orbits whose ground track repeats over a selected period of time are said to have a *repeating ground track*. These orbits are often used when it is desired to collect measurements over certain locations on the surface of Earth at regular time intervals. The former Soviet Union designed a unique orbit for reconnaissance satellites referred to as the *Molniya* orbit (after the satellite of the same name). This orbit is characterized by its high eccentricity (0.7), 12-hour period, and critical inclination of 63.4° . The latter allows fixing apoapse over the former Soviet Union to permit long periods of communication.

Orbital Maneuvers

The subject of orbital maneuvers could be a separate article unto itself; however, we can provide an introduction here. We focus our discussion on the Hohmann transfer, the most common technique for transferring a satellite from one orbit to another.

Hohmann Transfer. Hohmann (14) proposed that the minimum-energy transfer between two coplanar orbits could be achieved using tangential burns, where the direction of the velocity change is tangential to the orbit. Hohmann considered only circular orbits, but the theory can be extended to elliptical orbits when the burns are applied at periapse or apoapse (where the flight path angle is zero). Figure 12 demonstrates the geometry of the *Hohmann transfer* for these two cases. If r_{initial} is the radius of the lower orbit (or the periapse radius for the elliptical case) and r_{final} is the radius of the higher orbit (or the apoapse radius of the final elliptical orbit), the semimajor axis of the transfer orbit and the

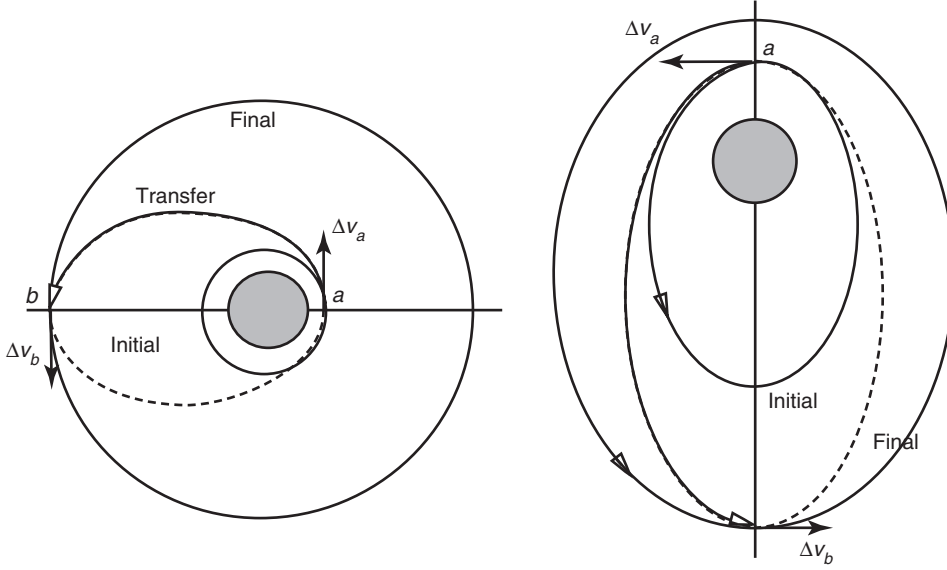


Figure 12. The Hohmann transfer (Fig. 5-4, Reference 7).

transfer time can be readily computed:

$$a_{\text{trans}} = \frac{r_{\text{initial}} + r_{\text{final}}}{2},$$

$$T_{\text{trans}} = \frac{P_{\text{trans}}}{2} = \pi \sqrt{\frac{a_{\text{trans}}^3}{\mu}}. \quad (74)$$

The magnitude of the two velocity changes can then be computed as

$$\Delta v_1 = v_{\text{trans}}^a - v_{\text{initial}},$$

$$\Delta v_2 = v_{\text{final}} - v_{\text{trans}}^b,$$

$$v_{\text{initial}} = \sqrt{\frac{2\mu}{r_{\text{initial}}} - \frac{\mu}{a_{\text{initial}}}},$$

$$v_{\text{trans}}^a = \sqrt{\frac{2\mu}{r_{\text{initial}}} - \frac{\mu}{a_{\text{trans}}}},$$

$$v_{\text{final}} = \sqrt{\frac{2\mu}{r_{\text{final}}} - \frac{\mu}{a_{\text{final}}}},$$

$$v_{\text{trans}}^b = \sqrt{\frac{2\mu}{r_{\text{final}}} - \frac{\mu}{a_{\text{trans}}}}. \quad (75)$$

A common application of the Hohmann transfer is transferring communications satellites from the initial low Earth orbits, into which they were launched, to

their final geostationary orbits from which they operate. Another common application is for interplanetary navigation (e.g., Earth to Mars).

Orbit Determination

A detailed discussion of the subject of orbit determination is outside the scope of this book. See References 15–17 for a thorough treatment of this topic. However, the concept that we have available an initial position and velocity for the satellite (or equivalently a set of Keplerian elements) is fundamental to much of this article, and thus a brief discussion of the topic is warranted.

The process of orbit determination can be described as developing a dynamic model for a satellite orbit, including its position and velocity at some reference epoch, that best agrees (usually in some least squares sense) with space- and ground-based tracking observations of the satellite. Tracking measurements can be provided in a variety of forms [azimuth/elevation, range, range-rate (Doppler)] and can be made from Earth-fixed tracking stations or from other satellites such as the Global Positioning System (GPS) (18) or the Tracking and Data Relay Satellite System (TDRSS). The tracking observations tend to be nonlinear functions of the satellite position, and the dynamic models that describe the motion of the satellite are nonlinear as well. Thus, the basic approach to orbit determination involves linearizing the measurement and dynamic models, numerically integrating the equations of motion for the satellite using a guess of the initial position/velocity, and then adjusting the state vector (which contains the initial position/velocity of the satellite, drag parameters, gravity parameters, empirical parameters, and tracking station coordinates) to provide the best agreement in a least squares sense between the tracking observations and the numerically integrated orbit. Such a dynamic technique is usually necessary because of gaps in the tracking observations, although new continuous methods such as GPS are allowing implementation of more kinematic methods.

Summary

This article has presented the basic theory for the motion of Earth-orbiting satellites. Although the assumption of two-body dynamics is quite limiting, this theory forms the basic tools used in astrodynamics today. When coupled with general or special perturbation techniques, a high degree of accuracy can be obtained. Indeed, orbit determination accuracies for some Earth satellite missions have approached the few centimeter level (19), as demanded by the science applications.

ACKNOWLEDGMENTS

The author thanks David A. Vallado for permission to use figures from his book.

BIBLIOGRAPHY

1. Battin, R.H. *An Introduction to the Mathematics and Methods of Astrodynamics*. Dover, New York, 1987.

2. Chobotov, V. (ed.). *Orbital Mechanics*. American Institute of Aeronautics and Astronautics, Inc., Reston, VA, 1996.
3. Danby, J.M.A. *Fundamentals of Celestial Mechanics*. Willmann-Bell, Richmond, 1992.
4. Prussing, J., and B. Conway. *Orbit Mechanics*. Oxford University Press, 1993.
5. Roy, A.E. *Orbital Motion*. Wiley, New York, 1988.
6. Szebehely, V., and H. Mark. *Adventures in Celestial Mechanics*, 2nd ed. Wiley, New York, 1998.
7. Vallado, D. *Fundamentals of Astrodynamics and Applications*. McGraw Hill, New York, 1997.
8. Nerem, R.S., C. Jekeli, and W.M. Kaula. Gravity Field Determination and Characteristics: Retrospective and Prospective. *J. Geophys. Res.* 100 (B8): 15053–15074 (1995).
9. Kaula, W.M. *Theory of Satellite Geodesy*. Blaisdell, Waltham, MA, 1966.
10. Jacchia, L.G. *New Static Models for the Thermosphere and Exosphere with Empirical Temperature Profiles*. SAO Special Report No. 332, Cambridge, MA, 1971.
11. Barlier, F., et al. A thermospheric model based on satellite drag data. *Annales de Geophys.* 34 (1): 9–24 (1978).
12. Hedin, A.E. MSIS-86 thermospheric density model. *J. Geophys. Res.* 92: 4649–4662 (1987).
13. Szebehely, V. *Theory of Orbits*. Academic Press, New York, 1967.
14. Hohmann, W. Die Erreichbarkeit der Himmel skorper (The Attainability of Heavenly Bodies), NASA Technical Translation TTF 44, Nov. 1960, Washington, DC, 1925.
15. Escobal, P.R. *Methods of Orbit Determination*. Wiley, New York, 1965.
16. Bierman, G. *Factorization Methods for Discrete Sequential Estimation*. Academic press, New York, 1977.
17. Tapley, B.D., B.E. Schutz, and G.H. Born. *Statistical Orbit Determination*. Academic Press, in press.
18. Hoffmann-Wellenhof, B., H. Lichtenegger, and J. Collins. *GPS Theory and Practice*. Springer-Verlag, New York, 1997.
19. Tapley, B.D., J.C. Ries, G.W. Davis, R.J. Eanes, B.E. Schutz, C.K. Shum, M.M. Watkins, J.A. Marshall, R.S. Nerem, B.H. Putney, S.M. Klosko, S.B. Luthcke, D.E. Pavlis, R.G. Williamson, and N.P. Zelensky. Precision orbit determination for TOPEX/POSEIDON. *J. Geophys. Res.* 99 (C12): 24,383–24,404 (1994)

R. STEVEN NEREM

Department of Aerospace Engineering Sciences
Colorado Center for Astrodynamics Research
University of Colorado
Boulder, Colorado

EARTH-ORBITING SATELLITES, DATA RECEIVING AND HANDLING FACILITIES

Ground support of spacecraft is often referred to as “TT&C,” for tracking, telemetry and command. Tracking generally refers to measuring the position of a

spacecraft, telemetry to information carried on the radio-frequency (RF) downlink signal, and command to information transmitted from the ground to a spacecraft. Increasingly, these terms are somewhat narrow to describe the action fully, as, for example, transmission of voice and video on the uplink. Though these functions were initially stand-alone elements of space operations, technological advances and economic pressures have blurred distinctions, both among them and with related operational activities. For example, computing power and reliability have advanced sufficiently that many functions can be economically accomplished at the point of ground receipt or of application, rather than at a centralized facility. For Apollo, the separate functions of command, telemetry, tracking, and voice communications were integrated into a single radio-frequency (RF) system. Readers should be aware that the boundaries of data receiving and handling functions are dynamic.

The genesis of data receiving and handling facilities was primarily in the military launch ranges. Differences and limitations necessitated changes as space activities developed. Geographically, coverage expanded globally. Operationally, mission durations became more continuous and ground activities increasingly interactive.

Early space flights were brief. Operations were simple, few, and far between. With no existing infrastructure, supporting systems, procedures, and operators had to be put in place for each mission—usually unique for each flight. Preplanning an entire mission has gradually been replaced by continual interaction by both flight controllers and mission scientists—resulting in far more productive missions. Led by the interplanetary designs, spacecraft have become much more autonomous—more robust and require less, and less critical, ground control. Some ground operations systems evolved from “specialized mission unique” to “general purpose multimission.” Even as multimission applications shared hardware, they often retained unique software. Generally, RF systems became more standardized (strongly prompted by international spectrum allocations), as have terrestrial communications as these needs are increasingly met commercially. However, many data systems and operating procedures remained quite unique. Sharing multipurpose assets is itself quite expensive, and technology has reduced costs sufficiently that in many cases it now makes economic sense to consider dedicated systems again—though constructed of standardized components.

Although much of this discussion focuses on NASA, it is in a larger sense the history of most space operations and thus has broader significance. In the Cold War environment the Soviet systems evolved independently, but most other agencies and countries followed NASA's lead, primarily to take advantage of existing designs, systems, and infrastructure. NASA originally had three separate networks of ground stations for controlling and tracking spacecraft. The Space Tracking And Data Network (STADAN) initially spanned the globe north and south across the Americas. It was used primarily for near-Earth scientific satellites. Telemetry and command systems were independent of each other, and tracking relied primarily on radio interferometry. The Deep Space Network (DSN) was used for interplanetary spacecraft; its notable characteristic was large antennae and very sensitive receivers. These stations were located around the world so that all directions in the celestial sphere were continuously in view of one of the stations. The Manned Space Flight Network (MSFN) was

established just for “manned” missions. These stations were located primarily along the ground track of early orbits launched from Florida.

By the late 1960s, NASA had about 30 ground stations around the world. In Spain, Australia, and California, one station of each of the three networks was located within a few kilometers of the other two—each independently operated and often having incompatible systems. Responding to budget reductions after the Apollo program, the agency consolidated the STADAN and MSFN facilities into the Space Tracking and Data Network (STDN). This consolidation was primarily in name, and years passed before it evolved into an integrated network. Systems and RF spectrum differences accounted for much of this delay. The ground stations slowly became more standardized, but the control centers frequently preserved their individuality. With the advent of relay satellites (see TDRSS Relay Satellites section following), NASA closed most ground stations and colocated remaining STDN facilities with the three DSN stations. The Tidbinbilla station (Fig. 1) near Canberra, Australia, is typical; the other two stations are located in California’s Mojave Desert and near Madrid, Spain.

These tracking stations were originally the focal points for executing space operations. Prepass planning was accomplished by the flight projects and the plan communicated—often via teletype messages on HF radio—to the stations. For the Mercury and Gemini missions, flight controllers were dispatched to the MSFN stations and interacted with the astronauts during the brief contact periods. As ground communications became sufficiently reliable, control was centralized in Houston for Apollo; the ground stations became remote communications and tracking terminals. The STADAN and DSN underwent similar evolution; control became real time and centralized at mission control centers as reliable communications became available.

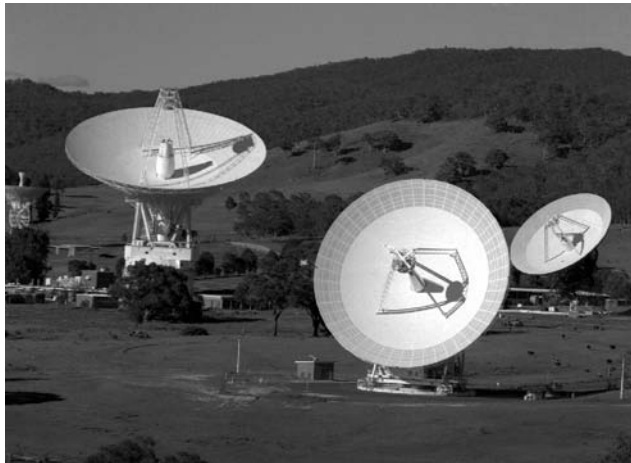


Figure 1. NASA’s Tidbinbilla, Australia, Tracking Station is one of three that provide communication with most NASA spacecraft beyond geosynchronous orbit. Antennae are (from left) a 26-meter STDN dish built for Apollo at Honeysuckle Creek and relocated to Tidbinbilla; the 70-meter diameter DSN antenna (expanded from 64 meters); the newest, a 34-meter beam waveguide DSN dish; and the original DSN antenna, a 26-meter dish expanded to 34 meters.

The subsequent evolution of data receiving and handling methods was synergistically enabled by advances in technology, although arguably more influenced by competitive factors than by either technical performance or economics. As often happens early in the evolution of a technology, there was little reason to standardize, and various designers developed somewhat different approaches to the same basic tasks. In space operations, the spacecraft manufacturer usually determined the design of the telemetric system. Industrial participants, of course, sought competitive advantages by perpetuating proprietary concepts. Internecine battles both between and even within space agencies also inhibited standardization. Ground facilities were usually viewed as infrastructure and were funded separately from the spacecraft programs. Thus space-ground cost trades were seldom a major consideration. As a result, ground facilities must often accommodate a number of different approaches to command and telemetry. This has led to such situations as one NASA Center now using four different ways to transmit telemetric data to the ground: time division multiplex (TDM) and three different packet data formats—each requires singular ground data-handling systems.

The economic benefits of standardization are persistently apparent, and progress is slowly being made. The Consultative Committee for Space Data Systems (CCSDS) was established by the major space agencies in 1982 to promote international standardization of space data systems. In 1990, CCSDS entered into a cooperative arrangement with the International Standards Organization (ISO), and many CCSDS recommendations have now been adopted as ISO standards. The reader is referred to the CCSDS for the latest information (<http://ccsds.org>).

Tracking

Initially, “tracking” was accomplished primarily by measuring the angle of the spacecraft’s path as a function of time from ground observing stations. Both optical and RF observations were used. Optical was by tracking telescopes, and, of course, depended on favorable lighting conditions. RF tracking was passive—receiving a spacecraft beacon signal at fixed antennae on the ground and determining angles using interferometry. NASA’s STADAN system was called the Minitrack. Radars were also used—transmitting signals from ground-based antennae and observing the reflections from the satellite. Radars provided another important parameter: range. Doppler information could also provide range rate data. Primarily due to weight and power limitations, few satellites carried transponders, and radars had to rely on skin tracking. The tracking data parameters measured by most of these methods were then transferred to central facilities with highly skilled specialists and usually the latest mainframe computer systems for processing and reduction. Except for a very few critical activities such as Apollo maneuvers, orbit determination was far from a real-time process. The tracking tasks included determining where the spacecraft had been when observations were made and also predicting its future orbital path. This propagation activity requires quite sophisticated modeling of such parameters as Earth’s shape and mass distribution, solar pressures, and outgassing and

venting from the spacecraft. The Apollo USB system measured range and range rate using a pseudorandom code carried on the uplink to the spacecraft, then turned around and retransmitted to the ground on the downlink. Some programs chose a simple tone rather than a pseudorandom code. Angles were also measured but were increasingly of secondary importance relative to the range–range rate data.

Spacecraft designers are beginning to incorporate GPS receivers onboard, permitting the spacecraft to determine its own orbit and transmit the data to the ground via telemetry. One of the more precise “tracking” examples to date has been the Topex-Posidon spacecraft, where GPS was used in the differential mode to measure the spacecraft orbit to about 2 centimeters. This precision was necessary because the spacecraft orbit was the reference from which changes in sea level were determined. Corner reflective mirrors are carried on some spacecraft, permitting range to be measured quite accurately by using ground-based lasers. Many of the applications of this technique were not to determine the spacecraft orbit, but rather to measure tectonic drifts on Earth—movement of the ground station.

Differences among various concepts include different protocols for time-tagging tracking parameters. Even within the same agency, some programs chose to read all parameters at the same instant, whereas others chose to read each parameter sequentially, each had its individual time tag. Such differences were necessarily reflected in different systems for ground processing. Today much of the activity in orbit computations has been transformed from art to science, mostly executed in software. At the same time, the tools—computers—have become cheap and reliable. Commercial vendors now offer software packages that determine orbit functions using low-cost computer platforms. This, along with the use of GPS, has greatly simplified the tracking function.

Telemetry

Telemetry—literally, measuring at a distance—is the transmission of data from a spacecraft to ground facilities. This often consists of measured parameters (either analog or digital) and could relate to either the status of the spacecraft or to observations made by the spacecraft scientific instruments (often referred to as “science data”). Occasionally this telemetry is simply relayed from another source via the spacecraft, for example, emergency locator beacons from downed aircraft and meteorological measurements collected from buoys. Voice communications from people onboard a spacecraft are often handled as “telemetry.”

Telemetric systems have been used in aeronautical and space flight testing for more than 50 years. Early telemetric systems were based on analog signal transmission. These systems used frequency multiplexed/frequency modulation (FM/FM) techniques. In this technique, each sensor channel is transmitted by mixing the sensor output with a local oscillator using dual-sideband suppressed-carrier (DSBSC) amplitude modulation (AM). The outputs of each mixer are then summed and used as the input to a frequency modulator for transmission. In this technique, each of the local oscillators forming the frequency multiplexed signal need to be different and separated in frequency to keep the channel signals from overlapping in the frequency domain.

The FM/FM techniques were used in the Mercury manned spacecraft program as well as in the early, unmanned programs such as the Ranger and Lunar Orbiter spacecraft. The need for a large number of oscillators and the lack of flexibility gave way rapidly to digital techniques. For example, the Gemini program began using digital techniques after the Mercury program, and this continued on to Apollo and the Shuttle (the major exception was the real-time video data that has remained in their TV-compatible analog format). Deep space missions to the Moon and planets have also been using digital techniques for sensor data since the mid-1960s. These systems are based on pulse code modulation (PCM) for the data acquisition. The modulation techniques can be either analog frequency modulation (FM) or phase modulation (PM) or digital techniques based on phase shift keying (PSK) or frequency shift keying (FSK).

In the following sections, we will examine the basics of the data acquisition system and data encoding. Then we will look at the digital modulation techniques and finally at the error-correcting codes used in current space systems. The International Foundation for Telemetry is a nonprofit organization formed in 1964 to promote telemetry as a profession. Interested readers are referred to <http://www.telemetry.org> for further information.

Data Acquisition. Sensor data is collected within the spacecraft, and the sensor output is typically in an analog form such as a voltage or a current. The analog value is converted to a digital code word by an analog-to-digital circuit in a process called pulse code modulation. The code word takes on a value from 0 through $2^N - 1$ where N is the code word size. Typical values for N range from 8 to 16 depending upon the degree of resolution required. The sample rate is determined by the bandwidth of the underlying analog signal. By the Nyquist sampling theorem, the sensor needs to be sampled at a rate at least twice the signal bandwidth, although sample rates at least five times the signal bandwidth are often used to allow for higher quality signal reconstruction and additional signal processing. In PCM systems, one needs also to specify the data encoding method and the time multiplexing format. Data encoding determines the channel transmission characteristics, and the multiplexing format determines the signal sampling discipline.

Data Encoding. The data format for transmission represents a trade-off among occupied bandwidth, data clock recovery, and immunity to phase reversals. As part of the overall telemetric system specification, the data encoding methodology needs to be specified. Typical data formats used in space telemetry systems include

- NRZ-L: Different voltage levels are assigned for logic 0 and logic 1 values, and the voltage level does not change during the bit period.
- NRZ-M: The voltage level remains constant during the bit period, and a level change from the previous level occurs if the current data value is a logic 1; no change is made if the current data value is a logic 0.
- NRZ-S: The voltage level remains constant during the bit period, and a level change from the previous level occurs if the current data value is a logic 0; no change is made if the current data value is a logic 1.

- Bi- ϕ - L: The NRZ-L signal is modulo-2 added with the data clock to show the logic level during the first half of the bit period and its complement during the second half of the bit period.
- Bi- ϕ - M: A level change is made at the start of the bit period, and a second change is made for the second half of the bit period if the logic level is a 1.
- Bi- ϕ - S: A level change is made at the start of the bit period, and a second change is made for the second half of the bit period if the logic level is a 0.

The M and S representations are called differential waveforms and are generally considered immune to phase reversals for correct decoding. However, the NRZ-M and NRZ-S techniques suffer from a higher error rate because when a single bit error occurs during transmission, two bit errors occur upon detection. Despite this higher error rate, the phase reversal immunity is considered a greater advantage, and differential NRZ encoding is frequently used. The L representations require correct phase determination to recover the data correctly. The Bi- ϕ representations generally require twice the transmission bandwidth of the NRZ representations. However, they have better clocking capabilities. Because of this, typical space telemetry transponders will have specifications detailing the maximum number of consecutive 0 or 1 values and the minimum number of transitions in a block when NRZ coding is used.

Data Packaging. For transmission from the spacecraft to the ground station, the data needs to be packaged to allow for synchronization, error detection, and accounting of data sets. Additionally, data may need to be tagged for routing to specific analysis centers. To accomplish this, data are typically packaged into one of two formats: telemetric frames and telemetric packets.

Telemetric frames are based on time-division multiplexing of the spacecraft sensor data for transmission across the space channel. Each sensor is transmitted at least once per frame cycle; more critical data is transmitted at a higher rate. The telemetric frame is based on a highly structured format that has embedded synchronization codes and accounting information to allow quick synchronization and verification of the transmission of all data. Normally, the frames are sent continuously to keep the channel filled to maintain synchronization and detect link dropouts. If link dropouts occur, the repeating, structured pattern of the frames allows rapid receiver resynchronization once the signal returns. The IRIG-106 standard developed for missile and aeronautical telemetry is often used in the space channel as well. The frames are processed in software upon reception using a database that shows the location of each sensor value in the overall frame structure and applies any necessary calibration to convert the raw PCM code value back to a meaningful datum.

As link reliability increases, a different philosophy for data packaging can be considered. Computer networking technology developed for ground telecommunications networks has been used as the basis for developing packet telemetric systems. The advantage of the packet system over the frame-based system is that the transmission can be more easily customized to the needs of the data sources, not sending data merely to maintain a fixed sampling discipline. The packets also allow distributed data processing.

The Consultative Committee for Space Data Systems (CCSDS) packet format is based on a structure consisting of a header that contains data source and destination information and other related system information. The body of the packet contains the data. Following the data, error detection, correction codes, and other information can be added. This packet system has been successfully used on many spacecraft, and commercial vendors supply standard ground station hardware for processing CCSDS packet telemetry.

Modulation. Once the telemetric data are packaged in a frame or packet format, they are transmitted across the radio link using some form of modulation. There are a large number of permutations for radio modulation depending upon mission needs, power requirements, and data rates. In this section, we examine the basic technologies used in data transmission.

The PCM/FM and PCM/PM techniques are often used in missiles and sounding rockets even today. The technique first uses the PCM technique to sample the data. The data are encoded using one of the data formats described and are placed in telemetric frames. Finally, the baseband data is routed to the input of either a frequency modulator (FM) or phase modulator (PM) for transmission. The advantage of using FM and PM is the constant envelope signal on the output.

The PCM/PSK/PM technique is an extension of the PCM/PM technique. The PCM output is used as the input to a phase shift keying (PSK) modulator. The PSK process used is binary PSK (BPSK) modulation which is a DSBSC AM process described by

$$s(t) = \sqrt{E}d(t) \cos(\omega t), \quad (1)$$

where \sqrt{E} is the normalized signal strength, $d(t)$ is the PCM output data as a function of time, and ω is the PSK subcarrier frequency. This technique has been used, for example, in the Voyager telemetric system. The additional PM modulation step provides two advantages. First, the technique allows mixing other subcarriers with the PSK data stream to provide a total composite signal. This technique also maintains a constant transmission envelope that can be important when transmitting through nonlinear amplifiers.

The use of the PM modulator does add a degree of system complexity, so that data transmission at the PSK modulator output level is frequently used. In the PSK system, many different combinations give different efficiencies (bandwidth and performance). These PSK techniques are supported in the standard TDRSS transponder hardware.

Binary-phase shift keying (BPSK) was described before. BPSK transmits one PCM bit per channel symbol and is the simplest of the PSK techniques. It also has the lowest transmission efficiency.

Quadrature-phase shift keying (QPSK) can be considered as two orthogonal BPSK channels transmitted at the same time. The transmitted signal takes two of the PCM bits at a time and forms a single channel symbol, where the channel symbol changes state every two bit periods. The two channels are called the “in phase” or I channel and the “quadrature phase” or Q channel. There is a total of four possible carrier phase states in QPSK, where the carrier is described by

$$s(t) = \sqrt{E_i}d_i(t) \cos(\omega t) + \sqrt{E_q}d_q(t) \sin(\omega t), \quad (2)$$

where $\sqrt{E_j}$ are the normalized signal strengths, $d_i(t)$ and $d_q(t)$ are the PCM output data as a function of time, and ω is the PSK carrier frequency. In the TDRSS system, the relative I and Q channel signal strengths can be varied; typical values are 1:1 for equal strength and 1:4 for nonequal strength. These are defined by the user profile when using the TDRSS system.

The QPSK modulation technique allows the possibility that the channel symbol state can cause a 180° phase shift. This can cause undesired harmonic generation in nonlinear amplifiers. To mitigate against this, a delay is added to one of the two bits used to form the channel symbol that keeps phase changes to a maximum of 90° . This technique is called offset QPSK (OQPSK). It is also known as staggered QPSK (SQPSK). The transmitted signal is then described by

$$s(t) = \sqrt{E_i}d(t + T)_i \cos(\omega t) + \sqrt{E_q}d_q(t) \sin(\omega t), \quad (3)$$

and the variables are as described for QPSK. The delay of one bit period T is seen added to the in-phase data channel. Upon reception, a similar delay is added to the quadrature-phase channel to bring the two bits back into alignment.

In many payloads, there are two independent PCM data streams: one for health and welfare data and one for high-rate scientific data. The TDRSS system allows considering the I and Q channels as two independent data channels that have different data rates on each. This form of transmission is called unbalanced QPSK (UQPSK). This technique has the advantage that it does not send the low-rate channel at the higher data rates, thereby degrading performance. The system also has the advantage of OQPSK that it does not have frequent 180° phase shifts. The mathematical description for the carrier is the same as that given before for QPSK.

By applying waveform shaping to the I and Q data signals used in the QSPK modulation signal, a form of frequency shift keying is made. The Advanced Communications Technology Satellite used the minimum shift keying (MSK) technique. The carrier can be described by

$$s(t) = d(t)_i \cos\left(\frac{\pi t}{2T_b}\right) \cos(\omega t) + d_q(t) \sin\left(\frac{\pi t}{2T_b}\right) \sin(\omega t), \quad (4)$$

where T_b is the bit period. The data filtering functions are half-period sinusoids. This technique can be demodulated by using a coherent QPSK demodulator.

The wireless communications industry is using another form of filtered QPSK called Gaussian minimum shift keying (GMSK). Instead of sinusoidal filter functions, Gaussian-shaped filters are used to prefilter the data before mixing with the carrier. The MSK techniques are being examined because of their transmission bandwidth efficiency compared with QPSK. A related, patented form of filtered QPSK developed by Feher is currently being evaluated for telemetric systems. This FQPSK modulation system uses proprietary filter functions in a manner similar to GMSK. Preliminary results indicate that it has an improved spectral efficiency for data transmission. Tests are currently being conducted within NASA and the U.S. DoD to evaluate this technique for ground and space telemetric systems.

Data Transmission. Several techniques are used for transmitting this modulated RF signal from a spacecraft to the ground. The original, and still most

common, method is direct transmission to receiving antennae on the ground. These transmissions may be initiated by ground command or instructions stored onboard or may be broadcast continuously. Signals are usually directed to specific TT&C stations, although transmission of some types of meteorological data, for example, is simply broadcast to all interested users. For weak signals, ground antennae are occasionally arrayed to provide increased gain.

Another technique increasingly used is relay through another satellite, usually one in geosynchronous orbit (see TDRSS section following). This technique involves a space-to-space cross-link in addition to the relay-sat-to-ground link, but it offers several advantages. Communications with Earth-orbiting satellites are possible almost any time, any place. (Depending on specific geometry, some limitations may exist in polar regions.) This avoids the need for a large network of ground stations. Such ground networks are usually expensive to build and operate and are subject to disruption due to local political and environmental conditions. It also centralizes the point of data reception on the ground, avoiding the need for subsequent terrestrial relay. Although cross-links can present some technological challenges, they also offer freedom from interference and interception.

Like the evolution of tracking data handling described before—and for similar reasons—telemetric data is usually transferred from the ground receiving station to a central facility for processing. Even though the telemetric data rates of today are greatly increased, economical hardware exists to perform much of this processing in a “black box” in real time at the ground station. However, as will be discussed under Future Trends, other factors limit this evolution.

The characteristics of the transmission medium and path are different in the space-to-ground link and in the terrestrial link from the ground station back to the control and central data handling centers. Although the latter was sometimes HF radio in very early days, it is normally satellite or coax/fiber today. These differences reflect directly that the “best” protocols for the terrestrial link are different from those for the space-to-ground link. Thus, the ground receiving station is required either to reformat the data, or commonly, encapsulate the first protocol into the second, increasing overhead.

As the Internet has grown, considerable interest is developing in using standard Internet data protocols (Transmission Control Protocol/Internet Protocol, TCP/IP) in spacecraft. This is still a research area, but there is considerable interest among scientific investigators and systems designers in providing Internet-to-space support. The ground-based TCP/IP protocols have some recognized difficulties in the space environment. For example, the long link delays and relatively high channel error rates lower the throughput of these protocols. To mitigate against these effects, research is currently ongoing into modifications of TCP/IP for space. Progress is also being made within CCSDS to develop the Space Communications Protocol System (SCPS) that is based on the TCP/IP protocol stack but modified for use in the space channel environment.

As part of the long-term trend for space communications, system designers are investigating the application of commercial telecommunications technology to space communications. This emphasis comes from two drivers: the need to reduce system cost and the need to use system bandwidth better. By using commercial telecommunications satellites and associated ground networks, spacecraft operating costs can be reduced because infrastructure costs may be

shared with commercial users. The need to use more efficient modulation techniques has led system designers to consider techniques used in the wireless industry. Modulation techniques such as GMSK and FQPSK were designed for the commercial wireless communications industry. The FQPSK technique is currently being evaluated within the U.S. Department of Defense for use in missile and aeronautical telemetry. Their relatively higher bandwidth efficiency over PSK methods makes them good candidates for use in constrained-bandwidth environments for space as well.

Data Processing. Telemetry was originally processed in decommutators that were essentially giant patch panels, reconfigured for each spacecraft. As computing power grew, processing was accomplished using software in mainframe computers. (Like IBM, many space pioneers found it difficult to advance beyond this stage.) The next step was use of application-specific integrated circuit (ASIC) technology. As ASIC technologies became practical, they proved much more efficient for repetitive processes such as telemetric processing than general purpose machines controlled by software. Also, processing in chips rather than mainframes permitted substantial increases in throughput telemetric rates, avoiding the need for expensive buffering to reduce processing rates. (They also proved considerably more reliable because most operating failures were in the connections.) The current state of the art is to use programmable gate arrays—combining the speed of a hardware approach with the flexibility of a software solution. Numerous vendors now offer such products.

Processing of data is often broken down into four levels: Level 0 through Level 3. Though the Level 0 step is still common, it is primarily an artifact of reel type onboard data recorders. Data were recorded serially onboard onto a recorder for playback (dump) to a ground station when one was in view. This data was then usually dumped with a reverse playback—avoiding the need to rewind and cue the recorder. The Level 0 function is to remove the space-to-ground transmission protocols, restore the reversed data stream to its original direction, and separate the various multiplexed parameters into their original components. Level 1 data processing involves appending axillary data from other sources, such as orbit and attitude data. Levels 2 and 3 processing require active involvement and judgment of the scientists in analyzing and presenting the data. Levels 1 through 3 are normally considered part of scientific data processing rather than the ground data-handling function.

The solid-state recorders increasingly used on spacecraft today, of course, do not need to be dumped in the reverse direction. Second, data can be read out of the recorder in any desired sequence, obviating the need to demultiplex a serial data stream on the ground. Third, using economical hardware and relatively simple software, space-to-ground protocols can be stripped and data reformatted for terrestrial transmission in realtime, even at 100-Mbps rates. This effectively obviates the need for Level 0 processing. Using GPS for onboard orbit determination, as described earlier, similarly obviates this Level 1 step.

Error-Correction Coding

Error-correcting coding has been studied by theorists since Shannon's work in the late 1940s. Error-correcting coding techniques became widely used in space

communications in the 1970s with the advent of good techniques by Viterbi and Fano for decoding convolutional codes. Current data services used with NASA's Deep Space Network and the TDRSS network permit the user to select from standard error-correction coding techniques.

Forward Error-Correcting Codes. Transmission over the space channel introduces errors into the data stream. The goal for space-to-ground transmission accuracy for most telemetric data has historically been a bit error rate (BER) of 10^{-5} , or no more than one error in 100,000 bits. (For comparison, coax and fiber generally deliver data at BERs better than 10^{-10} .) Two basic techniques are used for recovery: request a retransmission of the data, or add an error-correcting code to the data. On missions that have long propagative delays, requesting retransmissions often reduces the data throughput to unacceptably low levels. Thus the error-correction technique is usually preferred because it requires less memory and processing capabilities on the spacecraft.

Here, we discuss the two common error-correcting coding techniques used in space systems. They permit correction of the data upon reception, so they are called forward error-correcting (FEC) techniques. They accomplish this task by adding redundant information to the transmitted data. This implies that a higher transmission rate is needed and consequently more transmission bandwidth to achieve the same raw data throughput. This is considered acceptable when achieving a much higher data quality. The FEC can permit the system to operate as if it had a 3- to 6-dB higher received signal-to-noise ratio that it actually has—a significant benefit.

The FEC that has the longest history in space communications has been convolutional coding. This has been used on missions since the 1970s and even for deep space missions like Voyager. The convolutional encoder is built around a fixed-length shift register whose length is known as the constraint length K . Once each bit time, a new data bit enters the shift register, and multiple outputs are computed as a function of the current shift register contents. Typically two or three outputs are generated for each bit time, giving rise to a rate one-half or rate one-third coding. A Viterbi decoder is used to decode the data and correct for transmission errors. NASA requires that all S-band services on TDRSS use a rate one-half convolutional encoder. Typically the constraint length is $K=7$. TDRSS also supports the rate one-third coding, but it is not part of the CCSDS recommendations.

The Reed–Solomon coding technique is a form of block coding where the data stream is partitioned into fixed-length blocks and redundant information is added to the end of the block to form the transmitted signal. Reed–Solomon block coding is characterized by the number of information symbols and the number of code symbols in a transmitted block. The CCSDS recommends Reed–Solomon encoding using transmitted blocks of 255 total symbols that have 223 information symbols in the block. This is written as a (255,223) R–S code.

The CCSDS standards recommend using Reed–Solomon codes either by themselves or in conjunction with convolutional codes, as discussed in the following section. The TDRSS does not support Reed–Solomon decoding as a basic service but will pass encoded data to the user for decoding. Reed–Solomon codes can be used without convolutional coding through TDRSS if Ku-band services are being used, but not in the S-band services.

Concatenated Coding. It is known that the output of the Viterbi decoder produces errors in a burst pattern. To compensate for this, missions can use a concatenated coding technique. First, a Reed–Solomon code is applied to the data stream to form what is known as the “outer code.” These coded data are then encoded with a convolutional or “inner code.” The convolutional code removes the bulk of the channel transmission errors, and the Reed–Solomon code removes the burst errors from the convolutional decoder. The CCSDS recommendation for concatenated coding is to use a (255,223) R–S outer code with a rate one-half, $K=7$ convolutional inner code.

A recently discovered coding technique called turbo codes has begun to be seriously investigated for application in space communications. These codes were first published in 1993. They have the advantage of providing performance near the Shannon limit for channel performance, even at low signal-to-noise ratios. Recent experiments have shown cases with TDRSS where the ground station receivers could not lock onto the signal in low SNR environments even though the symbol decoding was still functioning. Turbo codes are based on convolutional codes on the encoding side and are decoded using iterative solution techniques in the receiver. Current research involves determining optimal code structures and decoding techniques for use in the space channel.

Command

Command is essentially the reverse of telemetry: digital data is modulated in one of several ways on the uplink carrier. Basic differences are that data rates are much lower, and transmission integrity is usually more critical because these data normally control the spacecraft. The command system generally can be thought of as providing instructions for the onboard spacecraft systems to initiate some action. Two classes of command strategies are commonly used in modern spacecraft: single command words and command files. Many early spacecraft command systems were basically tone response systems, that is, when a given tone of a specific frequency was received by the spacecraft, a specific action was undertaken. Tone commands are still often found in rocket command destruct systems.

CCSDS has published standards used by many programs. Differences are primarily in the intermediate transmission path between the originating control facility and actual transmission from the RF transmitter. Some concepts transmit the command message directly from the control center in “real time,” whereas others stage it at intermediate points such as the remote ground station. In the latter case, it is error checked and stored at this intermediate location, awaiting a transmit instruction either as a function of time or on electronic or verbal instruction from the control center. A “command echo” is frequently used for verification: a receiver at the ground antenna detects the command on the RF link, which is returned to the control center for comparison with the intended command structure. The term “command” is often used to describe all signals modulated on the uplink signal; however, such uplinks can include both analog and digital voice communications with astronauts, and video uses are planned in the future.

The command word will typically look like a binary computer instruction that has a system address, command instruction number, and command data value. Sometimes the command will use a simple check sum for error detection. Other error check strategies that have been used in spacecraft include repeating the command bits twice and then performing a bit-wise check for agreement before executing. In all cases, the command received word would be echoed in the telemetric data stream as a specific telemetric word, so that the operator could check for command validity. The spacecraft would be configured for two possible modes of command acceptance: execute the command upon passing the received error check, or wait until a ground operator sends an execute command after verifying that the operational command was correctly received. In operations, routine commands would be executed immediately upon passing error validation. Commands that could be dangerous for spacecraft operations, for example, firing a thruster or exploding pyrotechnic bolts, would have the operator in the loop. The operator would wait for the configuration command to be correctly echoed in telemetry and then would send the execute command to initiate the actual operation.

Command files are more appropriate for sequences of operations, especially those for a spacecraft beyond Earth orbit. For example, the Voyager planetary encounters would operate by loading a file of command sequences for camera settings and spacecraft attitude before the encounter where the communication times could be in excess of hours. The operators would verify that the file had been loaded into the spacecraft memory by playing back the memory through the telemetric stream. Any transmission errors could be corrected before the encounter, and then the spacecraft would operate in automatic mode, without immediate operator interaction, during the encounter. The structure of the commands within the files can resemble those found in the individual command mode. The file organizes the commands for bulk transmission.

All of the modulation and coding techniques used for telemetric transmission apply to command transmission as well. In general, very sophisticated FEC is not done because of onboard processing limitations. However, parity checks or check sums are often used to identify command errors. For low Earth orbit (LEO) operations, repeating commands that are found in error is often sufficient.

TDRSS Relay Satellites

In the mid 1970s, NASA began developing a communications and tracking network in space—the Tracking and Data Relay Satellite System (TDRSS). The basic concept was demonstrated during the Apollo Soyuz Test Project mission using the Applications Technology Satellite ATS-6. By using the relay satellite concept, communications with a fixed ground station for LEO satellites can be increased from 5 to 10 minutes per contact at a fixed ground station to more than 85% of the total orbit time. (The TDRSS concept used a single ground station that had two active relay satellites located above the horizon of this station. This resulted in a “zone of exclusion” on the opposite side of Earth that had no coverage. This ZOE region was later covered by another relay satellite, using a

second ground station, to meet the needs of the CGRO satellite, and increased TDRSS coverage to 100%.)

TDRSS was first seen as just another station of the STDN, except that it was located in the sky. However, technical differences such as signal levels and Doppler rates made it clearly a very different kind of station. The primary motivation for building TDRSS was economic, more for cost avoidance than cost reduction. Costs of upgrading the STDN to meet Space Shuttle needs, it was estimated, would be more than the cost of developing and deploying a relay satellite system. A second consideration was political; operations at ground facilities on foreign territory were disrupted several times by political disputes. The first TDRSS satellite was launched in April 1983. It was stranded halfway to geosynchronous altitude by an explosion of its upper stage rocket, but by using its tiny maneuvering thrusters, eventually it reached its operational location over the Atlantic that summer. Launching the second satellite was the objective of the ill-fated Challenger mission in January 1986. Completing the TDRSS constellation was a high priority because so many new programs depended on it. Thus, the third satellite was aboard Discovery in September 1988 on the first Shuttle mission after Challenger. It was completely successful; four more spacecraft have since been successfully launched.

From technical and performance standpoints, TDRSS has proven outstanding. In addition to supporting users at standard S-band frequencies, TDRSS provides a high-frequency service in the Ku-band region and commercial service at C-band. (This commercial service was added by the spacecraft contractor to spread the costs of the satellites and thus reduce costs.) This complex system has demonstrated its ability to relay data at more than 400 megabits per second from user spacecraft virtually anywhere over Earth. Each relay satellite can communicate with two high-rate users simultaneously, and the system can handle an additional 20 low-rate users. This low-rate system—termed multiple access—was one of the first systems to successfully use code division multiple access (CDMA) techniques that permit multiple users to share the same radio spectrum and is now employed in some cellular phone systems. TDRSS provides more than six times the global coverage of the ground network it replaced, at data rates six times as high. It enables real-time contact virtually anytime, anywhere—an especially valuable feature of the relay satellite approach. It also has unanticipated applicability to small systems and has been successfully used with high-altitude balloons, with aircraft to relay real-time infrared images for forest fire fighting, and with handheld transmitters for voice and data.

A fixed price procurement approach was selected even though extensive development was needed, and NASA became directly responsible for launching the satellites. Thus, substantial government–industry cooperation was required in the face of a somewhat adversarial contractual relationship. Another significant challenge was that of dealing with two distinctly different cultures: a communications industry accustomed to a highly regulated environment and the engineering focused aerospace industry. The difficulties were more complex than can be summarized here; interested readers are referred to Aller, R.O. *Issues in NASA Program and Project Management*. From a program perspective, the TDRSS system was deployed late and cost more than original estimates. Even so, the cost of

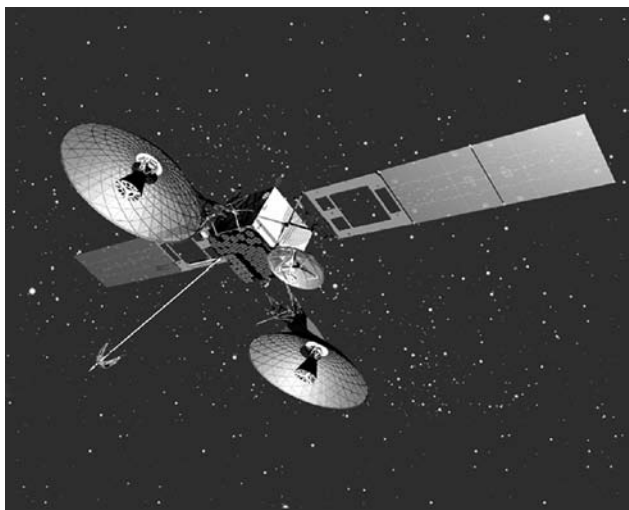


Figure 2. Second generation of NASA's TDRSS spacecraft, used for tracking and communications with most near-Earth-orbiting spacecraft. From geosynchronous orbit, the two large movable antennae each provide communication at high data rates with spacecraft far below. A phased array antenna on the Earth-facing panel permits simultaneous communications with 20 users at lower data rates. The small dish antenna relays signals to/from the ground control station in New Mexico.

augmenting the STDN for the Shuttle was avoided, and closing most of the STDN in the 1980s and shifting to TDRSS actually reduced costs. Amortized development and launch costs plus operating expenses of the TDRSS are less than the operating costs alone would have been for the ground network it replaced.

NASA procured three follow-on TDRSS spacecraft in the mid-1990s. Hughes (now Boeing) won the competition with a design based on its HS-601 bus. An artist's rendition of this design is shown in Fig. 2. The first of this series was successfully launched in 2000.

Radio-Frequency Spectrum Issues

By the nature of space operations, "wireless" communications links are necessary. Thus ground data-receiving facilities are strongly influenced by radio spectrum issues. The International Telecommunication Union (ITU) was established in 1865 (as the International Telegraph Union). The ITU is now under the United Nations and endeavors to coordinate radio spectrum use internationally. Radio waves have no respect for political boundaries, so this coordination has become critical. As transmitters and receivers moved from terrestrial locations into the air and then into space, signals radiate far more of Earth. As the private sector sought to capitalize on the economic potential of freedom from "wires" that technology subsequently enabled, extreme pressures have developed for spectrum allocations. This stage and the extremely rapid development of associated technologies presented an especially vital but difficult challenge to the somewhat cumbersome ITU processes. Not surprisingly, these pressures have focused on

the lower frequencies. Technology is more proven in this region, and signal characteristics such as power, directionality, and interference favor this region for widespread commercial use. Because wave-lengths are longer, hardware manufacturing tolerances are also less critical and costs are reduced.

For largely these same reasons, space activities initially operated in the lower frequency regions, primarily the VHF region. However, factors such as interference from other RF users in this region and the need for greater bandwidth slowly drove space programs to vacate this region in favor of higher frequencies. NASA chose the S-band region, transmitting uplinks near 2100 MHz using downlinks above 2200 MHz. Some other U.S. agencies elected to downlink at these same frequencies, but to transmit at lower frequencies (L-band, near 1700 MHz). This S-band spectrum was originally allocated internationally for space use in a footnote to the allocations table, specifically noting that such use was permitted on a secondary basis. This secondary allocation meant that use must not interfere with the use of this spectrum by those holding primary allocations and that conversely, holders of secondary allocations had no recourse if primary users interfered with them. Interestingly, this applied to programs such as NASA's Apollo and Space Shuttle programs: their primary operating frequency was authorized only as a secondary allocation, legally subject to these previously stated constraints. (Note that deployment of the TDRSS relay satellite had an interesting effect on this situation because TDRSS transmits downward to low Earth-orbiting satellites at the normal uplink frequency and conversely, receives signals transmitted upward at the downlink frequency.)

This intense competition for spectrum, together with need for additional bandwidth and desire for primary allocations, is driving space programs to adopt higher frequencies—mostly in the Ku and Ka-band regions. Initial use of optical frequencies is also being considered. Technologies for these higher frequencies is not as mature as those for lower frequencies, but they do offer advantages in bandwidth and in the power and weight of communications terminals in space. Disadvantages, in addition to less mature technologies, include the need for high pointing precision of antennae and susceptibility to atmospheric interference, especially rain and clouds. Certainly, costs of necessary modifications to the ground infrastructure inhibit these moves to higher frequencies.

Future Trends

As has been said, predictions are difficult, especially about the future! Tempting as it is to focus on the evolution of relevant technologies as the indicator of future trends, history suggests that sociological factors soon dominate technology and economics. Recall that for years after technology replaced the steam locomotive with diesel power, railroads retained firemen even though they no longer had boilers to stoke! Early TT&C facilities had the characteristics of "infrastructure," and infrastructures are especially adept at self-preservation, aided by the detachment from their customers. We will attempt to place data-receiving and data-handling facilities in the context of the historical evolution of technology and from this perspective identify factors determining trends. Finally, we will hazard some guesses about near-term trends.

Typically a technology evolves from an era of a few specialists proficient in using it, through one where the process becomes increasingly understood, standardized, and automated, so that necessary expertise and skill levels drop. Eventually tools become available that those unskilled in this particular technology can readily use to meet their needs. Entrepreneurs see market potential and derive new, more efficient products and processes by employing the technology. However, progress is frequently intermittent as other factors come into play. Those finding careers, organizations, and profits threatened summon political and industrial support to resist change, and efforts at standardization are soon caught up in competitive battlegrounds. Note that much of the early effort necessarily involved “inventing” the needed data-receiving and data-handling facilities. The entities formed for this purpose specialized in designing, rather than using, solutions. Once created, these entities usually strive to reinvent custom solutions for each application, rather than using or adapting existing alternatives.

Technologies that existed early in the space age determined the technical approaches for meeting satellite communications and data-handling requirements, and more importantly, they also determined the structure of the organizations and processes for meeting these requirements. Yet, technology seldom conveniently evolves within the boundaries of established management and fiscal structures. Matters that should be simple engineering issues whose solutions are submerged in silicone, for example, instead become contentious interorganizational issues. Few comprehend the technical implications as technology expands across borders, and interpersonal relationships—influenced by individual, provincial, and fiscal interests—dominate resolution.

The result of this confluence of technological, sociological, and political realities is that future trends are dominated by the heritage of the particular organization. New players usually adopt the latest economical technology and employ commercially available solutions when they enter the arena; thereafter they become increasingly bound by their own heritage. The pace and direction of technology limit advancement, but they are only two factors determining future directions. Implicit in this hypothesis is the view that few of the concepts and facilities commonly used today for receiving and handling space data from satellites are state of the art. Prudent conservatism is certainly warranted. Challenges of adopting state-of-the-art technologies are that performance risks are often increased due to immature development, costs are usually relatively high in the early stages, and amortization of changeover costs is difficult to justify. However, if a truly competitive marketplace develops using a given technology, this usually stimulates the private sector to evaluate objectively the economics of advances in that technology.

Although the expanding technologies of communications and computing are often said to be converging, semantics must be understood. Many of the recent advances in communications result from computing advances: the ability, effectively and economically, to digitize, package, route, and unpack material being moved. Transmitter and receiver advances, especially at higher RF frequencies as well as optical, contribute directly to space operations. (Other communications technology advances such as fiber are very significant, of course, but are not directly applicable to space-to-ground communications.) Considering the

computing aspect, approaches that combine the speed of a hardware approach with the flexibility of a software solution, as now offered by programmable gate arrays, will become dominant. Chip technology offers tremendous increases in processing speed over the old mainframe approach, due in part to the small dimensions.

Other prognostications: space programs will rely increasingly on navigation satellites (GPS, Glonass, Galileosat) to determine orbit onboard. A TCP/IP-like transmission protocol such as SCPS, adapting to the space channel environment, will probably become popular. Use of embedded chips at the points of application greatly streamline data processing, allowing eliminating some previously required no-value added steps such as transport to centralized facilities. Techniques such as turbo codes that promise to perform near the Shannon limit, even under low signal-to-noise ratio conditions, will be pursued. Constraints on spectrum and bandwidth will focus attention on more efficient modulation techniques, such as FQPSK. Spectrum competition and the need for additional bandwidth are also driving space programs to higher frequencies—mostly in the Ku- and Ka-band regions. Optical transmission will become a serious contender for the space-to-ground channel—circumventing RF spectrum restrictions, avoiding RF interference, and permitting increased security. Technology is improving the power limitations of light transmitting diodes, and spatial diversity (multiple receiving sites) will avoid cloud cover. Adoption of customized commercial hardware and software solutions in preference to unique solutions will greatly increase, especially among new users. Even noncommercial users will gravitate in this direction. Numerous, mostly small, vendors now offer space operations products commercially and competitively.

READING LIST

- Aller, R.O. Issues in NASA Program and Project Management, NASA SP-6101(02), 1989.
- Consultative Committee for Space Data Systems. Radio Frequency and Modulation Systems, Part 1, Earth Stations and Spacecraft, CCSDS 401.0-B. National Aeronautics and Space Administration, Washington, DC, 1994.
- Consultative Committee for Space Data Systems. Telemetry Summary of Concept and Rationale, CCSDS 100.0-G-1. National Aeronautics and Space Administration, Washington, DC, 1987.
- Feher, K. *Wireless Digital Communications Modulation and Spread Spectrum Applications*. Prentice-Hall, Upper Saddle River, NJ, 1995.
- Ferguson, C.H., and C.R. Morris. *Computer Wars: The Fall of IBM and the Future of Global Technology*. Times Books, New York, 1994.
- Griffin, M.D., and J.R. French. *Space Vehicle Design*. American Institute of Aeronautics and Astronautics, Washington, DC, 1991.
- Horan, S. *Introduction to PCM Telemetry Systems*. CRC Press, Boca Raton, FL, 1993.
- Jet Propulsion Laboratory. Deep Space Network/Flight Project Interface Design Handbook, 810-5, Revision D, Pasadena, California.
- Larson, W.J., and J.R. Wertz (eds.). *Space Mission Analysis and Design*, 2nd ed. Microcosm, Torrance, CA, 1992.
- National Aeronautics and Space Administration. Space Network (SN) Users' Guide, Revision 7. Goddard Space Flight Center, Greenbelt, MD, 1995.
- Postman, N. *Technopoly, The Surrender of Culture to Technology*. Vintage Press, 1993.

- Rappaport, B. *Wireless Communications Principles and Practice*. IEEE Press, Piscataway, NJ, 1996.
- Sklar, B. *Digital Communications Fundamentals and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1988.
- Telemetry Group Range Commanders Council. Telemetry Standards, IRIG 106-99. Range Commanders Council, White Sands Missile Range, NM, 1999.
- Yuen, J.H. (ed.). *Deep Space Telecommunications Engineering*. Plenum, New York, 1983.

CHARLES T. FORCE
Tracy's Landing, Maryland

STEPHEN HORAN
New Mexico State University
Las Cruces, New Mexico

EASTERN LAUNCH FACILITIES, KENNEDY SPACE CENTER

Introduction

The Eastern Test Range (ETR) is the familiar designation of a group of bases, facilities, and installations that support space vehicle launches from Cape Canaveral Air Station (CCAS) and the Kennedy Space Center (KSC) and ballistic missile test launches from or near the Cape Canaveral Air Station. It is operated by the United States Air Force's 45th Space Wing, which provides services to itself as a launch agency as well as to the other Government and private agencies that use the launch bases at Cape Canaveral and KSC. The expression "familiar designation" is used because the current official designation of the Range is simply the Eastern Range (ER). However, the term "ETR" is probably used more frequently than "ER." The Kennedy Space Center is the popular name of the National Aeronautics and Space Administration (NASA) launch installation on Merritt Island as well as the name of the NASA center that manages and operates it. The formal designation of the center is the John F. Kennedy Space Center, NASA.

Eastern Test Range

The ETR, or "The Range" as it is often called, extends from the Atlantic Coast of central Florida, including the well-known Cape Canaveral launch base and its Headquarters about 20 miles south at Patrick Air Force Base, southeast along the Bahamas, the West Indies, and on to Ascension Island in the South Atlantic Ocean. Recently, bases in Bermuda and Newfoundland have been added to support launches that have more northerly azimuths.

Patrick Air Force Base serves as the administrative center of the ETR and functions also as a data processing facility and as the location for some instrumentation such as radars and telemetry antennae that benefit from a different perspective of the vehicles in the early launch phase. Patrick also has an operating airfield that serves as a home base for both fixed-wing planes and helicopters supporting launches.

History. In 1946, after the end of World War II, a committee, established to determine the site of a long-range proving ground for missiles, chose Cape Canaveral, Florida, as the launch base. Actually, the Cape Canaveral site was the group's second choice. The preferred site was El Centro, California, whose range was down the Gulf of California along the coast of Baja California. This site had several advantages, including more predictable weather and the proximity to many of the American aerospace contractors. However, the government of Mexico refused to allow overflights of Baja California. Flight along the coasts of the Bahamas from Cape Canaveral was acceptable to the British government, which then controlled the Bahamas. A third candidate, in northern Washington State that had a range along the Aleutian Islands, was rejected for reasons of weather and remote location (1).

Cape Canaveral, even as a second choice, had many advantages, both technical and administrative. As shown in Fig. 1, the area was located on a promontory extending into the Atlantic Ocean and was not near any large population centers. It would allow launch over open water but still near enough to a number of islands that could support instrumentation for tracking the missiles.

In retrospect, the Cape Canaveral option proved exceptional in a way only visionaries at the time could have foreseen. When the Space Age dawned, the easterly launch azimuth presented the designers of space launch vehicle with an additional velocity contributed by the Earth's rotational velocity. El Centro whose launch azimuth was approximately south would not have had this

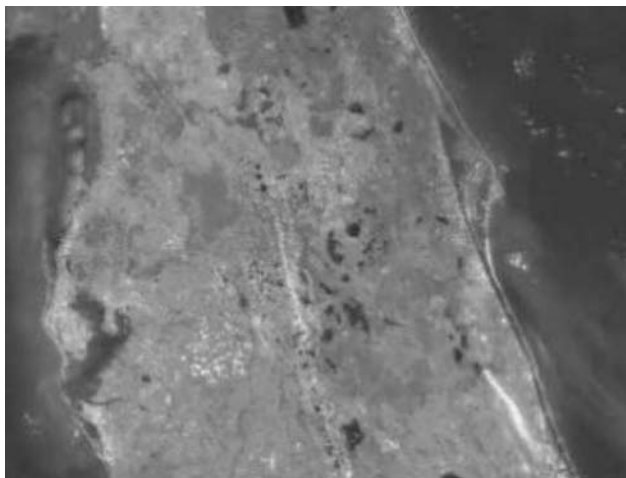


Figure 1. This map shows the location of Cape Canaveral on the east with respect to other Florida locations such as Tampa Bay on the west. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

advantage, and the Aleutian range that had a generally westerly azimuth would have had the penalty of having to provide additional velocity to compensate for the Earth's rotational velocity. A secondary advantage of Cape Canaveral was its lower latitude, a consideration that became important only after the advent of geosynchronous satellites. Of course, the eventual need for polar orbiting satellites would necessitate establishing a West Coast launch base that could safely accommodate southerly launch azimuths. Interestingly, when this need was recognized, the choice of such a base was not El Centro, but the area now known as Vandenberg Air Force Base in California.

The new range would be a joint project of the Air Force, the Army, and the Navy under the management of the Air Force. In the course of its history, there would be several changes of name, both of the Range itself and the military organizations managing it. The managing organizations included Advance Headquarters, Joint Long Range Proving Ground; Headquarters, Long Range Proving Ground Division; Headquarters, Air Force Missile Test Center (AFMTC); Headquarters, Air Force Eastern Test Range (AFETR); Detachment 1 SAMTEC, ETR; Headquarters, Eastern Space and Missile Center (ESMC); and finally in 1991, 45th Space Wing. The Installation would also undergo a number of different designations starting with Bahamas Long Range Proving Ground, Long Range Proving Ground, and proceeding through Florida Missile Test Range, Atlantic Missile Range, and Eastern Test Range to the present title of simply Eastern Range. There were good and sufficient reasons for the changing terminology, but the basic mission of the establishment remained substantially the same as always. Some variations were necessitated by the changes of clientele from aerodynamic missiles to ballistic missiles, and to the current situation where most of the customers are involved in launching spacecraft into Earth orbit.

After sufficient preliminary work had been accomplished, the new range was ready for its first launch, which, ironically, was not one of the aerodynamic missiles for which it had been designed. On 24 June 1950, a vehicle called Bumper, a two-stage vehicle consisting of a captured German V-2 as a first stage and a liquid propellant WAC Corporal as a second stage, was launched from Complex 3. The launch, historic though it was, was not a success; the vehicle was destroyed by Range Safety. Less than a week later, a second Bumper was launched but it was not a complete success (2). Although the launches did not meet their mission objectives, there were some positive results. The launch complex was validated, the tracking devices worked, and the capability of Range Safety to destroy a potentially dangerous missile was proven.

During the next several years, many different missile projects were launched and tested at the Cape. Early projects included Lark, Matador, Mace, Snark, Bomarc, and Navajo. The Eastern Range was also the launch and test site for the Redstone, Jupiter, Thor, Polaris, Poseidon, Trident, Atlas, Titan, and Minuteman missile systems.

The coming of space vehicles was signaled by the selection of the Vanguard Earth Satellite Project which was intended to launch at least one instrumented satellite into Earth orbit as part of the U.S. contribution to the International Geophysical Year in 1957 and 1958. The Navy was selected to manage this program based partly on the Naval Research Laboratory's experience with the

Viking sounding rocket and its associated scientific payloads. Although it was assigned to the Navy, it was an essentially civilian project that would not interfere with the Department of Defense military missile projects and did not carry the high priorities assigned to those projects. After the Soviet Union launched Sputniks 1 and 2 in October and November 1957, before Vanguard had attempted to launch a full-up vehicle, the Army was directed to proceed with preparations to launch a vehicle that would add a fourth stage to the Jupiter-C and attach to the stage a small scientific satellite. After the first Vanguard attempt to orbit a satellite failed in December 1957, the Army team launched the first U.S. satellite, Explorer 1, on 31 January 1958. Second attempts by both the Vanguard and Jupiter-C teams failed in February and March, before the first Vanguard success, Vanguard 1, on 17 March 1958.

The creation of NASA in October 1958 added a new presence to the Range. Immediately, the only discernible difference was the transfer of the Vanguard Project to the new agency, but soon a NASA Atlantic Missile Range Operations Office (AMROO) was established to represent NASA interests at the Range. The new NASA projects leaned heavily on the military IRBMs and ICBMs. The Delta (sometimes called Thor-Delta) was a further adaptation of the Vanguard upper stages that had the Thor as a first stage and was designed to place a variety of scientific and applications payloads into Earth orbit. NASA used a new Air Force launch vehicle, called the Atlas Agena that used the Atlas as a first stage and a new hypergolic propellant upper stage called Agena for many of its early lunar and planetary missions. The versatile Agena was also used as the upper stage for the Thor and the Titan at Vandenberg Air Force Base, but only Atlas Agena vehicles were launched from the ETR. The capability of the Atlas to place payloads into orbit without an upper stage allowed using it in NASA's first manned space program, the well-known Mercury project. Interestingly, the early space launches followed the pattern of the aerodynamic missiles and the succeeding ballistic missiles in suffering many failures, either of the launch vehicle or of the payload after a successful launch, in the early days of spaceflight.

The next logical step in spaceflight, after demonstration of the achievability and usefulness of space as a scientific and applications medium, was the introduction of the human into the space arena as a part of the mission aloft, rather than as a ground-based operator. The first of these to impact the Range was Project Mercury, a program designed to demonstrate that humans could successfully operate in space and be safely returned to Earth. Seven suborbital flights of spacecraft without astronauts aboard were conducted using both the Redstone and the Atlas as launchers between September 1959 and March 1961. Two successful flights and one unsuccessful flight were flown aboard Atlas, and three successes and one failure were experienced with the Redstone. One suborbital launch, Mercury Redstone 2 (MR-2), carried as a passenger a chimpanzee nicknamed "Ham" and resulted in a successful flight and recovery on 31 January 1961, the third anniversary of the launching of Explorer 1.

The Mercury and Gemini Programs. On 5 May 1961, Astronaut Alan Shepard was launched on a Mercury Redstone 3 rocket to a range of 260 nautical miles with subsequent successful recovery of the pilot and the spacecraft Freedom 7. On 21 July of the same year, Astronaut Virgil Grissom

made a similar flight aboard the spacecraft Liberty 7, although the spacecraft was not recovered.

Even before the Shepard and Grissom suborbital flights, an unsuccessful flight of an unmanned spacecraft on Mercury Atlas 3 occurred, followed by the successful flight of Mercury Atlas 4, also unmanned, on 13 September 1961. The next Mercury Atlas flight on 29 November 1961 carried the chimpanzee called "Enos" into a planned three-orbit flight with successful deorbiting and recovery of both the passenger and the spacecraft, although the flight was shortened to two orbits because of some flight anomalies. The program was now ready for the manned orbital phase that consisted of four flights of increasing duration. All resulted in successful flights and recoveries, though not without some anomalies and difficulties.

Astronaut John Glenn became America's first man in orbit aboard Friendship 7 launched into a three-orbit mission and successful recovery on 20 February 1962 by Mercury Atlas 6. Astronaut Scott Carpenter in Aurora 7 followed on 24 May 1962 on another three-orbit mission marred by an overshoot of the intended landing area by several hundred miles. On October 3, Astronaut Walter Schirra in his Sigma 7 spacecraft flew a successful six-orbit flight with subsequent successful landing and recovery. The Mercury program concluded on 16 May 1963, when Astronaut Gordon Cooper in Faith 7 spacecraft splashed down safely after a flight of 22 orbits of the earth, 34 hours and 20 minutes after lifting off from Complex 14 aboard Mercury Atlas 9 (3).

The next step in the manned programs was the Gemini Project, which would demonstrate the capability to rendezvous and dock in space, as well as amassing more experience on manned flight to achieve longer time in orbit and at higher altitudes. The program, as indicated from its name (the name of the constellation Gemini translates to The Twins), would use a two-person spacecraft and an Agena stage, called the target vehicle, modified to achieve multiple restarts and equipped with a docking adapter, by which the Gemini and the target vehicle would be physically attached. The Gemini were boosted by a modified Titan II launch vehicle from Complex 19. The target vehicles were boosted by Atlas Agena vehicles launched from Complex 14, the same complex that was the site of the orbital Mercury launches. On 14 April 1964, less than a year after the last Mercury launch, Gemini 1, an unmanned spacecraft, was launched successfully into orbit. On 19 January 1965, a suborbital mission called Gemini 2 was also successful. Between March and August of that year, three manned flights were successfully carried out in which the number of orbits was increased from three to 120 and the apogee was increased from 121 to 188 nautical miles. The next mission was to demonstrate actual rendezvous and docking with the target vehicle, but the Agena malfunctioned and did not achieve orbit. An alternate mission involving rendezvous of two Gemini spacecraft was substituted, and the program continued while the Agena problems were addressed and solved. The program, which ran through November 1966, ultimately achieved rendezvous on that mission and on five additional missions and achieved docking with the target vehicle on four. There were difficulties with the Gemini spacecraft itself on one mission and both the target vehicle and a substitute on another, but the final three missions were all successful in rendezvous and docking and included some extravehicular activity (EVA) on all three (4).

The Eastern Range Facilities

Cape Canaveral Air Station. The heart of the ER and its predecessors is, of course, the launching area itself; without it there would be no requirement for the remainder of the ER, but it would not carry out its mission without the rest of the ER. In its own history, it has undergone a number of name changes. Originally designated as Operating Sub-Division 1, it was redesignated as Cape Canaveral Auxiliary Air Force Base in 1951, Cape Canaveral Missile Test Annex in 1955, Cape Kennedy Air Force Station in 1964, Cape Canaveral Air Force Station in 1974, and Cape Canaveral Air Station (CCAS) in 1990. Throughout all of the nomenclature machinations, the basic mission did not change, and the area was familiarly known as “The Cape.”

Launch Complexes. If the *raison d'être* for the ETR is the Cape, then the reason for the existence of the CCAS is the collection of the launch complexes located there. Typically, a launch complex consists of a launch pad, a service structure, an umbilical mast or structure, and a launch control center.

The launch pad is the platform on which the missile rests during launch preparations and from which it is actually launched. This platform contains a multiplicity of connections to the vehicle. Fluid connections and the necessary controls are included to load propellants into the vehicle and to unload in case of a launch delay or at the conclusion of a servicing or “tanking” test. Pneumatic connections are included to allow loading gases used as pressurants or in small control devices. There are also electrical connections that allow controlling vehicle functions remotely during launch preparations and instrumentation connections for monitoring the state of the vehicle and pad systems.

The service structure, variously and familiarly known as a gantry or a tower, gives workers access to the vehicle during launch preparations and can be of assistance in positioning and assembling the vehicle on the launch pad. Typically this structure has a number of platforms that allow circumferential and vertical access to the assembled vehicle. Necessarily, the service structure must be removed to a safe distance for the actual launch, so many of these platforms must either fold or retract to allow movement of the structure. Most service structures are mounted on railroad type wheels that allow movement to the retired positions on permanent tracks. Some projects have used service structures that can be rotated from the vertical to the horizontal to the vertical for assembling the vehicle, serve as a series of work platforms for prelaunch activities, and then rotate from the vertical to the horizontal before final launch preparations.

Some functions, which for reasons of accessibility cannot be routed through the launch platform, must be maintained even after the service structure has been retired or rotated away from the vehicle. These are typically accomplished by a connection on a mast that remains near the vehicle even after the service structure is withdrawn. These are generally referred to as “umbilical” masts or towers. The connections are severed either by movement of the vehicle at liftoff or by a mechanism on the mast itself, which senses liftoff and initiates disconnect. Originally, umbilical masts were rather flimsily constructed and had to be replaced after each launch. Recent improvement in materials and design have made possible permanent umbilical towers that require minimum refurbishment after each launch.

The launch control center, commonly called the blockhouse, was generally a heavily reinforced structure from which most prelaunch launch operations were supervised and from which the launch crew directed the actual launch. It provided a safe environment for the launch crew to direct operations during final launch and during other hazardous operations.

Other installations on a launch complex might include propellant tanks, both cryogenic, such as liquid oxygen and liquid hydrogen, which must be stored at high pressure to maintain the liquid state; so-called storable propellants, such as hydrazine and nitric acid, that can be kept at ambient temperatures; the necessary fluid lines and controls to load and unload the vehicle; and pressurant tanks that typically contain liquid or gaseous nitrogen for maintaining so-called blanket pressures in tankage when empty.

The following is a listing of the launch complexes at the Cape and a description of the projects which used them, some of the majors events associated with them, and their current status. Complexes were given numerical designations, 1, 2, 3, etc. Generally, if a complex has two launch pads they were designated A and B although there were some exceptions to this practice. Some complexes were planned but never built. This accounts for the fact that some numbers are not listed. Most of the complexes that are still active or play a noteworthy role other than for launching are mentioned specifically.

Complexes 1 and 2 were constructed for the Snark winged missile program and also supported some launches of the Matador, another aerodynamic missile. These complexes remained active until 1960.

Complexes 3 and 4 were designed to support the interceptor missile Bomarc. Complex 3, however, did support the Cape's first launches—those of the Bumpers in April 1950. It was later used by other projects, including the X-17 used to test reentry nose cones, some early Redstones, and some Polaris operations.

Complex 5 and Complex 6 were used for the Redstone project, an Army ballistic missile whose range was approximately 200 miles. There were two launch pads with a common blockhouse, and the facility was usually called Complex 5/6 or simply 56. The complex is best known as the site of all six Mercury Redstone missions, including that of America's first man in space, Alan Shepard, and the subsequent flight of Virgil (Gus) Grissom. It was also used for several early Explorer and Pioneer missions, including the first successful American lunar flyby mission by Pioneer 4 launched by a Juno 11 on 3 March 1959.

Complexes 9 and 10 were used for the long-range, rocket-booted, ram jet powered aerodynamic missile Navaho.

Complex 11 was used for the Atlas ICBM. It was also the site of the first orbital use of the Atlas. On 18 December 1958, almost the entire Atlas, minus jettisoned booster section, was placed into orbit as Project SCORE. The vehicle was equipped with communications equipment which was used to broadcast President Eisenhower's Christmas greeting to the world from space.

Complex 12 was used for the numerous Atlas ICBM launches and for two unsuccessful launches of multistage Atlas Able vehicles on projected lunar missions in 1960. NASA's flights of the Ranger Program, designed to take a series of television views of the Moon before impact, were launched from here as were

several Mariner missions that flew by Venus and Mars. The first Orbiting Astronomical Observatory (OAO) and two Orbiting Geophysical Observatories (OGO) were launched from Complex 12. Some OGOS were launched from the Western Test Range into near circular polar orbits; they were nicknamed POGO for Polar Orbiting Geophysical Observatory to distinguish them from the ETR launched OGOS which were nicknamed EGO for Eccentric Orbiting Observatory. The first three Applications Technology Satellites were also launched from here.

Complex 13 was used by both the USAF and NASA for their Atlas Agena launches. The USAF launches were usually classified missions; NASA missions launched from Complex 13 included one OGO/EGO and five Lunar Orbiter missions that provided photographic coverage of most of the Moon's surface in support of the Apollo Program.

Complex 14 was used for the Atlas ICBM project and for the NASA Atlas Mercury Project. It was also used for an unsuccessful launch of a lunar mission by the Atlas Able vehicle in November 1959. The Atlas Agena target vehicles for the Gemini Program were launched from here.

Complexes 15 and 16 were used for the Titan ICBM program. Complex 16 was also used for the Pershing solid propellant missile.

Complex 17 was designed for the Thor Intermediate Range Ballistic Missile. It has also been used for the Thor-Able, Thor Able Star and Delta launches by the USAF, NASA, and Boeing (formerly McDonnell Douglas). The Boeing usage is for commercial and government contracted launches. Current vehicles still being launched at Complex 17 include both Delta 2 and Delta 3 vehicles. Complex 17 stands ahead of all other complexes on the Cape in the number of missile and space launches. From the first Thor launch in January 1957 through the end of 1998, the complex supported 276 launches. The number of first and spectacular missions is so large that only a few can be mentioned, with the obvious danger of omitting some equally outstanding in some opinions. From a technological point of view, the first restart of an upper stage flight in flight, that of the Able Star second stage on the launch of the Transit 1B Navy's navigation satellite on 13 April 1960, is important. TIROS 1 (for Television Infrared Observation Satellite), America's first weather observation satellite, was launched by a Thor Able on 1 April 1960. The active communications satellite Telestar 1, the world's first privately owned satellite was launched on a Delta on 10 July 1962. SYNCOMs 2 and 3 and Intelsat 1 are important milestones in the development of geosynchronous satellites. The Mars Pathfinder mission, with the Rover Sojourner, launched on 4 December 1996, is a more recent accomplishment. A 1991 aerial view of Complex 17 is shown in Fig. 2. This photo was taken from the east (Atlantic Ocean) side and shows the two service towers around two Delta vehicles being prepared for launch. The solid propellant strap-on motors are clearly visible on the vehicle on Pad 17B, to the viewer's left.

Complex 18, a two-pad complex, was built for the Thor IRBM program. When the Navy's Vanguard Earth Satellite project for the International Geophysical Year was established with a short lead time until its first launch in 1956, arrangements were made with the Air Force for the new program to share Complex 18. All 14 Vanguards were launched from Pad 18A, Thors were launched from 18B, and the blockhouse and other facilities were shared. The Air



Figure 2. Complex 17, which is still active, is shown in the 1991 photograph, taken from the east. Note the Delta vehicle on the left pad, 17B. Even though the service structure surrounds the vehicle, the thrust augmenting boosters are visible. The wedge-shaped blockhouse is located at the intersection of the covered cable trenches running west from each pad (USAF photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Force later used the complex for launches of the Blue Scout, a version of NASA's four-stage solid propellant launch vehicle.

Complexes 19 and 20 were built for the Titan ICBM projects. Complex 19 was also the launch site of the NASA Gemini two-person spacecraft. Complex 20 was modified to accommodate the Titan IIIA, which consisted of the two liquid stages of the upcoming heavy-lift IIIC space launch vehicle. Four successful launches of the Titan IIIA were successfully conducted by May 1965, after which the complex was deactivated.

Complexes 21 and 22 were built for the USAF Bull Goose decoy missile program and were also used by the Matador follow-on program called MACE.

Complexes 25 and 29 were built for the Navy's Fleet Ballistic Missile Program. Launches of Polaris, Poseidon, and Trident I took place here.

Complex 26, a two-pad complex, was built for the Army's Jupiter IBM and Army/NASA Juno space programs. It was the site of the launch of the first U.S. Earth satellite, Explorer I, on 31 January 1958. The successful flight and recovery of the two space monkeys, Able and Miss Baker began from Complex 26 on 29 May 1959 (5). Complex 26 was declared a national historic landmark in 1964.

Complex 30 was a dual launch pad used for the Army's solid propellant Pershing missile program.

Complexes 31 and 32 were built for the USAF solid propellant multistage Minuteman ICBM. They included both surface and below ground (silo) launch capability.

Complexes 34 and 37 were built for NASA's Saturn 1 and 1B heavy-lift space vehicles. They were the sites of five vehicle development launches of these vehicles, some of which carried payloads as secondary objectives. Complex 34 was the site of two suborbital launches of Apollo hardware, as well as the launch Apollo 7, the first manned orbital flight in the program. It also has the undeniable distinction of being the site of the fatal spacecraft fire that took the lives of Astronauts Grissom, White, and Chaffee. Complex 37 saw the orbital launches of five unmanned tests of Apollo hardware. Complex 37 has been assigned to Boeing under the Air Force contract for its Evolved Expendable Launch Vehicle (EELV).

Complex 36, built for the Atlas Centaur space vehicle, has two pads, 36A and 36B. The first stage of the rocket is a modified Atlas, and the second stage (Centaur) is the first successful cryogenic (hydrogen/oxygen) upper stage. It has been used by NASA for a number of space launches, including lunar and planetary missions. It is still in regular use for USAF missions and by the Lockheed Martin Corporation (formerly General Dynamics) for commercial launches and for NASA contracted launches. The Surveyor soft landing missions to the Moon in the early 1960s; the interplanetary Pioneer 10 and 11 missions to Jupiter, Saturn, and eventually out of the solar system; Mariner missions to Mercury, Venus, and Mars, as well as a number of commercial and government communications satellites were among the many launches from Complex 36.

Complex 39 is at the Kennedy Space Center. It was designed for the Apollo program's Saturn V lunar launch vehicle and was also used for the Saturn 1B vehicles used in the Skylab and Apollo-Soyuz Test Program launches. It has been modified for Space Shuttle requirements and is still in regular use. It will be discussed further in the section of this article called The Kennedy Space Center.

Complexes 40 and 41 were built for the USAF's Titan IIIC space vehicle program; construction was completed in 1965. This approach, sometimes called ITL, for Integrate, Transfer, and Launch, was a new concept for the Cape. The core vehicle consisting of two Titan stages and upper stages were assembled and prepared in the Vertical Integration Building (VIB) on a platform on wheels that also served as the launch vehicle and launch umbilical tower. This part of the whole assembly was then transported by locomotives to the Solid Motor Assembly Building (SMAB) where the two large solid motor boosters, already checked out in the SMAB, were attached to the core. This assembly was further transported on the rails to the launch pad itself where the launch platform was secured in place. A service structure was provided at the pad to facilitate payload installation and final preparations for launch. The VIB was located several miles from both the SMAB and the launch pad, so it could be used as laboratory and office space without the personnel constraints of the traditional blockhouses or launch control centers. Complex 41 was modified to accept NASA's Centaur upper stage for the NASA Titan Centaur Viking Mars lander/orbiter program and Voyager outer planets program and the joint Germany/US Helios solar research program in the 1970s. Complexes 40 and 41 have been further modified to accept the larger USAF Titan 4 vehicle that can accommodate either a Centaur upper stage or the solid propellant Inertial Upper Stage (IUS) or can accomplish certain

missions without an upper stage above the Titan IV core vehicle. They are still used by the USAF for military missions and by Lockheed Martin for either commercial launches or for NASA spacecraft launches. Complex 41 will be used by Lockheed Martin under the Air Force contract for its Evolved Expendable Launch Vehicle (EELV) after the Air Force's need for the complex for Titan IV launches. Some other facilities in the ITL and elsewhere on the Cape have been assigned to either Lockheed Martin or to Boeing for their EELV programs.

Complex 46 was built between 1984 and 1986 to support the Navy's Trident II ballistic missile program. Between January 1987 and 1989, 19 Trident II launches were conducted from this complex. With no new Trident requirements for the complex forthcoming, Spaceport Florida Authority (SFA), with the aid of an Air Force grant, redesigned the complex to accommodate small commercial space launch operations. The completed complex has been used by Lockheed Martin for launches of variants of its Athena space launch vehicles for such missions as NASA's Lunar Prospector and the commercial/international ROCSAT for the Republic of China Satellite (6).

Industrial Area and Missile Assembly Buildings (Hangars). The launch complexes are the heart of the launch base, but they could not exist without a myriad of other functions and facilities at Cape Canaveral. Many of the functions are provided in the Industrial Area, which is located a safe distance from the launch complexes. These include such prosaic things as cafeterias, dispensaries, office buildings, distribution of utilities such as water, electricity, communications and sewage, maintenance of roads and grounds, security, fire protection, and janitorial services to give a few examples. Less prosaic features of the Industrial Area are the Missile Assembly Buildings as well as spacecraft and space payload preparation buildings.

An obvious need exists for a facility to receive the launch vehicle from the factory, inspect and refurbish it, and prepare it for transfer to the launch complex for launch, as well as to store missiles not yet ready for flight or awaiting a pad assignment. These functions are accomplished in Missile Assembly Buildings commonly referred to simply as hangars. In the early hurry-up days of the missile programs, a standardized structure, essentially an aircraft hangar, was provided by the Range to incoming projects. The use of a standardized building expedited the availability of the hangars for the new projects which were typically operating under high priority and accelerated schedules. The typical hangar had an area of about 41,000 square feet. A high bay area with a number of cranes was included along with large sliding doors at both ends of the hangar. On each side of the high bay was a two-story concrete block structure. This could be used by the project for offices, guidance laboratories, telemetry and data labs, etc. The office and lab spaces were generally air conditioned because of the hot and moist coastal Florida climate; the high bay areas themselves were not air conditioned; environmental control for the high bay consisted of opening and closing the large doors.

Propellant Storage and Explosion Proof Areas. There are functions that are neither compatible with the industrial area nor is it feasible to locate them near the complexes where their work may be interrupted by launch activity. Such facilities include propellant storage areas and explosion proof areas. Those launch complexes using propellants such as RP-1, a kerosene-like substance used

as a fuel in many launch vehicles; liquid oxygen, a cryogenic oxidizer used in conjunction with RP-1; liquid hydrogen, a cryogenic fuel used in some space vehicle stages; and other commodities such as liquid and gaseous nitrogen and liquid helium maintained storage vessels in the complexes for direct servicing of the vehicle during countdown. It is, however, the responsibility of the ETR to maintain sufficient quantities of these commodities to keep the storage tanks at operational levels. At one time, the ETR actually operated a LOX (familiar name for liquid oxygen) and LIN (familiar name for liquid nitrogen) manufacturing plant on Cape Canaveral. They are now supplied commercially, but the ETR is responsible for maintaining supplies and facilities to fill project needs.

Other projects use storable, that is, noncryogenic, liquid propellants sometimes as either the principal propellants for the missile stages or as the propellants for upper stages or spacecraft. The principal fuel of this type is a hydrazine variety, and the usual oxidizer is a form of nitric acid. For some projects, these propellants can be loaded into the vehicle well in advance of the actual launch countdown, so that there is no need for last minute loading capability at the complex as that for cryogenic propellants. Other projects prefer to load these propellants during the launch countdown and maintain on-pad facilities to maintain strict temperature control and to measure the load precisely. The ETR maintains storage facilities for the commodities and delivers them to the complexes as needed.

Solid propellant launch vehicles such as Minuteman and Pershing included storage facilities for these motors in their launch complexes. However, the need in space vehicles for solid propellant upper stages for actual spacecraft injection into orbit dictates that a facility for storing such motors be provided until the motors enter the actual processing flow which is normally when the solid motor is mated to the spacecraft. The use of solid propellant strap-on motors in the first stages of space launch vehicles to increase performance has resulted in the need for storage facilities for many such items until they enter the launch preparation flow. The ETR operates such storage facilities for the various Range users.

The advent of sophisticated satellites that have their own propulsion and controls system dictated the need for explosion proof areas where satellites could be fueled and their control systems serviced. Other such facilities accommodated the mating of a spacecraft to a solid propellant upper stage and sometimes a dynamic balancing of a spin-stabilized upper stage/satellite combination. Safety concerns dictated locating such facilities in relatively remote areas of the Cape.

Range Control Center. As the blockhouse or launch control center is the focal point for the actual launches, so is the Range Control Center the focal point for all of the functions that the ETR performs to support a project, as well as the center of its own mission of providing Range Safety. Here, all the information is gathered that is necessary for the ETR to ascertain that it is ready to perform its functions for a launch. Ground safety and security must verify that the launch danger area is clear; instrumentation coordinators must ascertain that the necessary radars, telemetry sites, and other instrumentation necessary for either the project requirements and safety monitoring are ready; and range safety officers must verify that there are no ships or aircraft in the prohibited areas and that the instrumentation required for the range safety function is operational.

These operations are coordinated by such officials as the Superintendent of Range Operations (SRO), Range Control Officer (RCO), and the Range Safety Officer (RSO). In earlier days, this building was in the heart of the industrial area, but it has been replaced by a new facility near the south end of the Cape.

A historical footnote to the reporting of status by instrumentation sites involves the famous expression "AOK" which is now generally attributed to the manned space program. Actually, this term has been in use at least at the ETR for many years before the manned space programs. When asked for readiness to support over communications circuits from remote sites, the operators at the sites had a choice of three replies. AOK meant that the site was operational and that it would support its requirements. CNY meant that the site was not operational and did not expect to be operational for the operation: BEX meant that the site could not meet its requirements at the time of the report but that it anticipated that it would be operational by the time its support was scheduled. The use of three letters for each readiness condition was intended to assist in recognizing the status signal over potentially noisy communications circuits.

Skid Strip. The landing strip is another prominent feature of the Cape. It was built in the early 1950s for the Navaho X-10 test vehicle. It was used in the mid 1950s as a landing area for Snark missiles which might be reused after successfully landing. The Snarks were equipped with landing skids instead of wheels; hence the name skid strip. Originally a 10,000 × 200 foot configuration, it has since been widened to 300 feet, resurfaced, and expanded to include a taxiway and a parking apron (7). It is frequently used for cargo aircraft delivering launch vehicles and spacecraft to the launch site. The skid strip has also been used by passenger aircraft bringing distinguished visitors to the USAF and NASA areas for tours or for observing important launches. The skid strip can be seen clearly in Fig. 3, an aerial view of the Cape.

Downrange and Tracking Stations. As early as 1954, the Eastern Test Range had three tracking sites equipped with tracking, telemetry, or photographic instrumentation. There was launch coverage instrumentation at the launch site at Cape Canaveral, and there were additional stations at Jupiter Auxiliary Air Force Base south of Cape Canaveral on the Florida mainland and at Grand Bahama Island in the Bahamas. Additional stations usually called Auxiliary Air Force bases were soon added on the islands of Eleuthera, San Salvador, Mayaguana, and Grand Turk in the British West Indies; in the Dominican Republic on the island of Hispaniola, and at Mayaguez in Puerto Rico. A submarine cable was incrementally constructed connecting all of the stations and was completed in December 1956. Some instrumentation was also located at Patrick Air Force Base to afford different viewing angles than was possible from the Cape itself during the launch phase.

The distance covered by these stations (about 1200 miles from the Cape to Mayaguez) was sufficient for the early programs. However, when the much longer range aerodynamic Snark and Navaho missile programs were approved, much longer distances had to be accommodated. Antigua and St. Lucia in the Lesser Antilles and Ascension Island, approximately 5000 miles from the Cape in the South Atlantic, were selected, and negotiations with the British governments ensued. A fourth at Fernando de Noronha, an island off the coast of Brazil, was also selected, and an agreement with the Brazilian government was negotiated.



Figure 3. A 1961 aerial photograph of the Cape, with North at the top, showing the Skid Strip in the center, the various launch pads generally along the coast of the Cape, and the Industrial Area to the left of the west end of the Skid Strip (NASA/USAF photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Antigua, St. Lucia, and Ascension were operational in 1958, and Fernando, as it was usually called, became operational in the summer of 1958. Even with the additional range stations, there were areas, where air breathing vehicles at relatively low altitudes, could not be tracked by land-based instrumentation. As

many as 12 small telemetry ships were positioned downrange to cover the gaps between the island-based instrumentation. As a matter of history, the ETR supported its first 5000-mile long mission (a Snark test flight) in October 1957 (8).

Although the Snark and Navaho programs were the impetus for extending the range, the new Intermediate Range Ballistic Missiles (IRBMs) such as the Army's Jupiter and the Air Force's Thor, as well as the Intercontinental Range Ballistic Missile (ICBMs) such as the Atlas, Titan, and Minuteman programs, after a slow start, eventually became major range users. A few years later, the new space vehicles became the largest users in terms of launches, and the Navy's Trident program remains a major user. Ballistic missiles and space launch vehicles that have higher trajectories than aerodynamic missiles could be tracked continuously from fewer stations, allowing the phase out of some downrange stations as the workload shifted from aerodynamics to ballistic and space launches. There have also been some notable additions to "downrange." Additional ships that have more sophisticated instrumentation were added and subtracted as workloads varied. Instrumented aircraft were also added that operated from bases in South Africa and Mahe in the Indian Ocean.

The downrange stations were usually officially designated as Auxiliary Air Force Bases, but there was frequently only one military person on site, usually an Air Force Officer, who held the title of Base Commander and among whose many duties was dealing as United States Government official with the local authorities. In the early days of Range operations, there was a Base Manager, an employee of the Range contractor (for many years Pan American World Airways) who ran the day-to-day operations of the station except for the instrumentation. A person called the Instrumentation Manager was an employee of Pan Am's instrumentation subcontractor, the Radio Corporation of America (RCA) Service Company. Most of the personnel at a downrange station were either from the Air Force, Pan Am, or RCA, although indigenous personnel were employed where feasible. For simplicity, the terms Pan Am and RCA are used here rather than trying to trace the various contracting and subcontracting schemes used during the history of the Range.

Range Instrumentation. The Range used a variety of instrumentation in its host responsibility of providing data to the Range users and in its own responsibility of ensuring that the launched vehicles did not pose a hazard to life or property. Trajectory information for both purposes was principally obtained by radars that tracked a transponder or beacon in the vehicle. These radars were also capable of echo or skin tracking, but the use of airborne beacons improved both the range to which something could be tracked and the accuracy of the data. There were other tracking devices used ranging from very complicated systems to such simple devices as optical gunsights and wire sky screens for close tracking, but precision tracking radars were the principal means for obtaining trajectory data. Telemetry antennas of varying sophistication were another important asset provided by the Range.

Command/control as used today is principally a method of either terminating the thrust of an errant vehicle or initiating an airborne destruct package in a vehicle if it has malfunctioned. This is the common conception of command/control, but there have been launch vehicle projects which have used this system to operate some vehicle systems usually as a backup to an airborne system.

Range Support and Safety. The preparation and launch of missiles, space launch vehicles, and spacecraft are potentially dangerous, so each participant in the process has a serious responsibility for safe conduct of such operations. The vehicle or spacecraft manufacturer or contractor, as well as the sponsoring agency, typically have safety organizations, but the Range as host has the overall responsibility for operations conducted on its facilities. To carry out this responsibility, the Range requires of its tenants detailed information about such hazardous items as pressure vessels, pyrotechnic devices, radioactive sources, and propellants that are used.

In the exercise of its responsibility to ensure flight safety, the Range requires each project to equip its vehicles with devices to assist real-time tracking by the Range, explosive or other means to either stop propellant flow (cutoff) or actually to destroy the vehicle (destruct), and receivers in the vehicle to accept cutoff and destruct radio-frequency signals from the ground and transmit these received signals to the appropriate devices. In perhaps overly simple terms, the Range Safety Officer (RSO) is presented with a plotting board display or display that shows the actual trajectory as measured by a tracking radar, for instance, and the predicted instantaneous impact point of a vehicle if thrust were terminated at a given instant. These are overlaid on a prepared chart showing such things as the nominal trajectory, the estimated normal deviation from the nominal, lines showing what areas are to be protected, and lines showing how far from the protected areas pieces of a destroyed vehicle might travel because of such factors as their size and density or human reaction time.

The RSO is also presented with other data, such as telemetry, indicating the performance of the vehicle. If the RSO determines that a flight is performing unsafely or that it will violate destruct criteria, the RSO can send appropriate signals to the vehicle which, when received by the command receivers aboard, will initiate cutoff and/or destruct (9).

Meteorological support is another important service of the Range. Even relatively routine information such as temperature, precipitation, and winds becomes vital when certain operations are scheduled. For instance, hoisting a spacecraft to the top of a launch vehicle by a crane cannot be done in high winds. The safety function of the Range itself demands certain ceiling and visibility minimums to allow optical tracking and observation of a launched vehicle early in flight. Weather balloon data on weather aloft is provided to the projects to allow determining whether such things as wind velocity and shear might compromise their control systems. Lightning data and prediction is important to both ground tasks and to launch operations. Activities on launch complexes are understandably restricted during thunderstorms. It is also necessary to know the state of the atmosphere aloft before a launch, so that a rising vehicle would not trigger lightning as it passed through dangerous levels.

In addition to the support given to launches from the ETR and KSC, the Range is called upon to support launches from ships and from submarines in the Navy's Polaris, Poseidon, and Trident ballistic missile programs. Support has also been rendered to United Kingdom submarine launches. Airborne launches of the Air Force's Hound Dog air-to-ground missile were a frequent customer between 1959 and 1965. Seventy-seven Hound Dogs were launched from B-52 aircraft over the Range in that period. Support was also given to Skybolt, the

proposed, but eventually canceled, successor to Hound Dog in the early 1960s. The Range has also given some support to the European Space Agency's Kourou launch site in French Guiana. Recently support has been given to the air-launched Pegasus space launch vehicle.

Kennedy Space Center

Introduction. The Kennedy Space Center (KSC) consists of almost 84,000 acres and an additional area of about 56,000 acres of submerged, formerly state-owned land. The location of the Kennedy Space Center with respect to Cape Canaveral is shown in Fig. 4.

Like the Eastern Test Range, KSC has undergone changes of name in both of its definitions. Organizationally, it was originally the Launch Operations Directorate (LOD) of the Marshall Space Flight Center (MSFC) in Huntsville, Alabama. MSFC was created in December 1958 when the Development Operations Division, including the Missile Firing Laboratory at Cape Canaveral, of the Army Ballistic Missile Agency at Redstone Arsenal in Alabama was transferred to NASA. In July 1962, LOD was raised to the status of an independent NASA Center and became the Launch Operations Center (LOC). After President Kennedy's death in 1963, the name was changed to the John F. Kennedy Space Center, NASA. This is usually shortened to Kennedy Space Center, or simply KSC. As a physical entity, the area was originally called the Merritt Island Launch Area (MILA); that eventually also became the Kennedy Space Center.

There were other NASA elements operating in the Cape Canaveral area in this period. The Navy's Vanguard Project was transferred to the Beltsville Space Flight Center (soon renamed Goddard Space Flight Center) of NASA when that new agency was created on 1 October 1958. The Cape-based Vanguard Operations Group (VOG) became the Field Projects Branch of GSFC and later the Goddard Launch Operations Division, which eventually assumed launch responsibility for Delta, Atlas Agena, and Atlas Centaur launches at the Cape, as well as for Thor Agena launches at Vandenberg Air Force Base in California.

The coming of the Mercury program to the Cape brought a detachment of the Space Task Group to Florida to support the Mercury Redstone and Mercury Atlas launches. This group eventually became the Florida Operations Group of the new Manned Spacecraft Center (MSC) in Houston. This was later renamed the Lyndon B. Johnson Space Center (JSC). This Florida Operations Group was transferred to the Kennedy Space Center in December 1964 after the conclusion of the Mercury program and before the first manned launch in the Gemini program.

The Atlantic Missile Range Operations Office which had been the official conduit for NASA's support requirements to the Range and the Range's commitment to support NASA was disbanded, and its functions were taken over by KSC's NASA Test Support Office at Patrick Air Force Base. The consolidation of NASA launch elements in Florida under KSC was finalized in October 1965 when the Goddard Launch Operations Division was transferred to KSC and became the Unmanned Launch Operations Directorate.

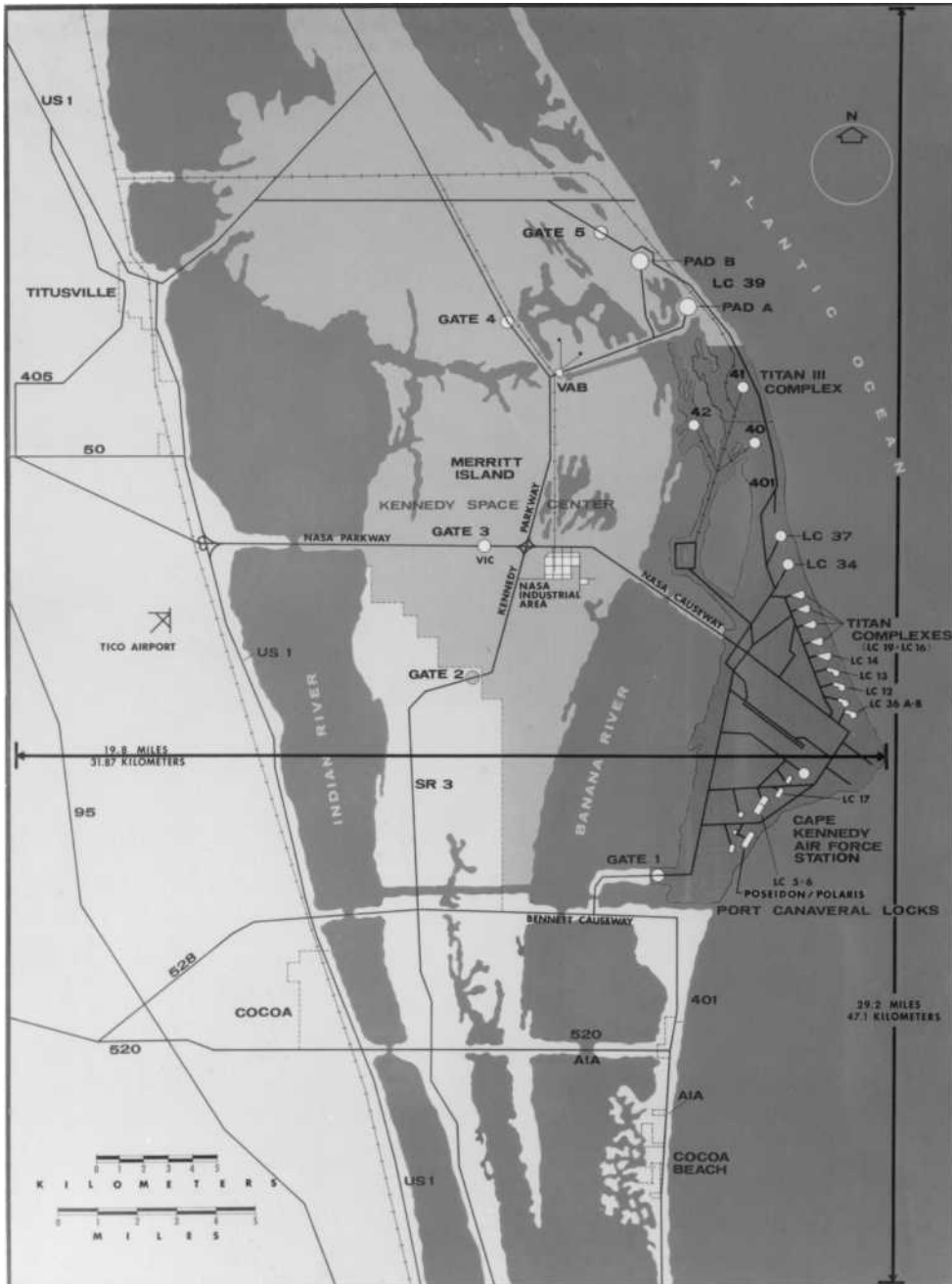


Figure 4. The geographical relationship of KSC, the Cape, and the surrounding communities in 1971. This drawing shows some facilities which had been planned but never built, such as Pad 42 (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The Apollo Program. The Apollo program was NASA's implementation of President Kennedy's declaration of 25 May 1961 that America "should commit itself to achieving the goal, before this decade is out, of landing a man on the Moon and returning him safely to the earth." At this time, there was no launch vehicle of sufficient power to do this, no spacecraft to take astronauts to the surface of the Moon and get them back to Earth, and no facility from which to launch the vehicle and spacecraft. The launch vehicle task fell largely to MSFC, and the spacecraft portion was assigned to MSC/JSC. The task of establishing a base for preparing and launching these elements clearly belonged to KSC.

Even though the launch vehicle, which would emerge as the Saturn V, had not been finally designed, it was clear that its size would preclude the possibility of using an existing Cape complex or of building a new complex on the Cape. A joint DOD/NASA task group pondered the question locating the launch base, and, after considering alternatives, including sites in Hawaii, Texas, California, Georgia, and islands in the Caribbean, selected Merritt Island. Other sites offered some appealing characteristics, but Merritt Island offered the opportunity to use many of the existing capabilities of the Atlantic Missile Range (AMR), and it allowed LOC/KSC to continue operating on the Cape at the same time as the Merritt Island Launch Area (MILA) was being built.

Complex 39 and the Vehicle Assembly Building. It was decided that the vehicle and its launch platform would be integrated vertically in a "hangar" on MILA to minimize time on the pad. The Apollo spacecraft elements would be checked out in the MILA industrial area and mated with the launch vehicle before the entire combination of launch vehicle, spacecraft, and launcher would be transported to the pad in the vertical position. There would be the capability of returning the combination to the hangar in case of weather or technical problems. The hangar, which would be larger than any existing one to accommodate the eventual 363-foot launch vehicle, is now known as the Vehicle Assembly Building (VAB). Its dimensions are 716 feet in length, 518 feet in width, and 526 feet in height. The VAB consisted of a high bay containing four bays or checkout cells, each large enough to accommodate a mobile launcher carrying a fully assembled space vehicle; a transfer aisle between the pairs of checkout cells; considerable office space, and a low bay for work on individual stages and components. A multiplicity of lifting devices as large as the two 250-ton bridge cranes with hook heights of 462 feet were included. Upward telescoping doors on each cell allowed access from and to the outside of the building. Only three of the high bay cells in the VAB were activated.

Operations in the VAB and at each of the two launch pads were controlled by personnel in the Launch Control Center (LCC) constructed to the southeast of the VAB. Not to be confused with most launch control centers or blockhouses on the Cape, this was a four-level building whose function was to check out the vehicle during preparations both in the VAB and at the launch pad. It was about three and a half miles from the pads, so it was not in the danger area and could provide support, such as offices, labs, data stations, and even a cafeteria, in addition to the actual control rooms called firing rooms. There were four of these firing rooms provided but, like the cells in the VAB, only three were fully equipped. A cable way and a personnel access corridor were provided between the third level of the LCC and the VAB.

The launch pads themselves were a major construction feat. Because the selected area was marshy, as was much of MILA, the entire pad had to be elevated with the bottom of the flame trench at ground level. Beneath the surface of the elevated pad were four floors containing terminal connection rooms, high-pressure gas systems, and emergency egress rooms for the astronaut crew and the pad close-out crew. A water system to cool the flame from the rocket engines was also there. Propellant storage tanks and other facilities similar, except in size, to Cape complexes for liquid fueled projects were also included within the pad perimeter. Pads 39A and 39B are essentially identical; several identical pads were planned at one time but were never built.

Essential to the mobile concept of operations were the mobile launchers, transportable steel structures that moved the Saturn V launch vehicles from the VAB to the launch pad. Three identical mobile launchers were built, each 445 feet tall and weighing 12,600,000 pounds when carrying an unfueled Apollo Saturn stack. The launcher consisted of a two-level base whose area is about a half acre and an umbilical tower capped by a 25-ton hammerhead crane. The bases contained a computer linked to counterparts in the firing rooms. The umbilical tower provided work platforms; swing arms that were disconnected automatically at liftoff; and distribution lines for propellant, pneumatic, electrical, and instrumentation functions. The highest swing arm, Number 9, 320 feet above the base connected with the Apollo spacecraft and was used by the astronauts to enter the command module. The base also provided hold-down arms that precluded motion of the vehicle after engine ignition until it was determined that all engines were operating satisfactorily. A rectangular opening in the launcher base allowed exhaust from the first stage engines to vent into a flame trench at the pad.

The launcher and the vehicle had to be moved from the VAB to the pad. After such concepts as barge transport, a rail system, and even pneumatic tired transporters and air cushion devices were considered and discarded, the choice was made of a crawler rolling on eight tracks that had a propulsion system similar to some used in strip mining. Two of these were procured. In addition to its transporting function, the crawlers could lift the entire stack to disengage it from the 22-foot pedestals in the VAB and lower it to similar pedestals at the pad after transport. The combined mass of the crawler, launcher, and unfueled Apollo Saturn was about 18 million pounds.

The highway for the crawlers to transverse with their burdens is the crawlerway, a two-lane road with a grass median to be straddled by the crawler. This highway is not paved but constructed of materials that can bear the intended weight without permanent deformation. The actual surface is Alabama river rock that minimized friction on the crawler treads. About halfway between the VAB and Pad 39A, an extension of the crawlerway turns northeast and goes to PAD 39B.

Another major element was the Mobile Service Structure (MSS), a 410-foot, 10,500,000-pound structure designed to give 360° access to the space vehicle and outfitted with air conditioning, elevators, computers as well as television instrumentation, communications, and power systems. It was removed from the pad by the crawler about 7 hours before liftoff and placed at its park site along the crawlerway. It was mobile and was removed from the danger area before launch, so only one mobile service structure was built to service both pads.

The size of the Saturn V and Apollo components required a new look at delivering them to the launch site. Earlier programs delivered either by road or by transport aircraft. The first and second stages were transported by barge from MSFC in Huntsville, Alabama. The third stage, the instrument unit, and the Apollo spacecraft, were delivered in a specially modified aircraft, called the Guppy.

Air transported components landed at the Cape's Skid Strip and were transported overland to the VAB. The barge method required constructing, on MILA, a dock and turning basin near the VAB, inclusion of draw spans on bridges, constructing a lock between Port Canaveral on the Atlantic Ocean and the Banana River, and considerable dredging.

The KSC Industrial Area. In addition to the Complex 39A area, there was a need for an industrial area for much of the same purposes as the Cape industrial area. One important difference was that there were no hangars in the KSC industrial area; essentially all work on the vehicles themselves was done at Complex 39. There still remained the requirement for administration and data gathering facilities as well as for a Flight Crew Training Building and for a building to prepare the command, service, and lunar modules. The Operations and Checkout (O&C) Building, a multistoried structure with a high bay checkout area 100 feet high and 234 feet long and an adjacent low bay checkout area 251 feet long was constructed to fulfill this last requirement. Salient features in the high bay were two altitude chambers each 50 feet high and 30 meters in diameter, which allowed checking out the Apollo spacecraft in near vacuum. The O&C also had a large and more conventional portion separate from the checkout area but connected by above-ground access corridors. The north portion of the O&C contained offices, laboratories, a weather station, and astronaut quarters. The first NASA occupants of the KSC Industrial Area were members of the MSC's Florida Operations Group that moved much of its operation from Hangar S on the Cape to the O&C building in the fall of 1964, even before that group was part of KSC. Completion of other buildings in the industrial area, including the Headquarters Building, the Central Instrumentation Facility, a Training Auditorium and the Base Operations building, followed soon after.

Apollo Operations. Even while all this preparation for the eventual Moon landings was going on at KSC, work on the project was proceeding on the Cape at Complexes 34 and 37. Orbital launches of boiler plate or dummy Command and Service Modules (CSMs) were conducted as early as 1964 by Saturn 1 vehicles and continued through July 1965. Suborbital tests of CSMs were conducted using Saturn 1B vehicles in 1966. The success of all of these tests led up to preparations for the first manned launch of a CSM from Complex 34 using the Saturn 1B vehicle AS (for Apollo Saturn) 204 and CSM 012. The 27 January 1967 fire in the spacecraft during a simulated countdown took the lives of Astronauts Grisom, White, and Chaffee. Modifications of hardware and procedures were made, which culminated in a successful orbital launch of SA 205 carrying a redesigned CSM 101 with astronauts Schirra, Eisele, and Cunningham. This was the last manned launch from the Cape, and Complexes 34 and 37 were deactivated.

The history of the Apollo Program with its successes, especially the accomplishment of the first manned lunar landing within President Kennedy's schedule and its difficulties, including the disastrous AS-204 spacecraft fire and the harrowing flight of Apollo 13, is well known and is covered elsewhere (10). The



Figure 5. The launch of Apollo 11, the first lunar landing mission, on 16 July 1969 (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

launch of Apollo 11 on 16 July 1969, which culminated in the first manned lunar landing and successful return to Earth, is shown in Fig. 5.

Skylab. The next challenge for KSC was Skylab, a program that would demonstrate that humans can exist in space for long periods and carry out useful scientific and applications tasks. (See the article on Skylab elsewhere in this Encyclopedia.) The first launch would orbit the Skylab itself, and three succeeding launches would propel crews of three men to the orbiting workshop where they would transfer to Skylab and live and work there for periods increasing from 28 to 59 to 84 days, before returning in their command and service modules. KSC facilities were modified for checkout of the workshop which replaced the Saturn third stage in the stack. The flights to the workshop did not require the power of the Saturn V; a Saturn 1B would suffice. At one time, it was planned to make these launches at Complex 37 on the Cape which had already supported the Apollo 7 manned launch in 1967. However, an ingenious plan to launch these missions (and the subsequent Apollo-Soyuz Test Program (ASTP) mission) from Complex 39 was devised and allowed deactivating Complex 37 years earlier than once planned at a considerable saving in money and manpower.

The solution was the construction of a steel pedestal, humorously called the “milk stool,” 127 feet in height atop which the much shorter Saturn 1B would be placed on a mobile launcher allowing some of the Saturn V and Apollo access arms to be used. The mission concept called for the launch of Saturn V with the workshop on one day and the launch of the Saturn 1B with the first Skylab mission crew on the next. This of course required that launch crews at KSC would have to prepare two launch vehicles at the same time and operate two complexes simultaneously. Plans were made to do this, and KSC was ready to launch the crew on 15 May 1973, the day after the workshop was launched. The experience in rendezvous and docking first demonstrated in Gemini and used extensively in Apollo also came into play here, as did the requirement for relatively short launch windows to accomplish rendezvous and docking. However, serious problems arose with the workshop during ascent, and it was necessary to delay the Saturn 1B launch while personnel at MSFC and JSC worked on ways to salvage the mission. These were quickly accomplished, and the Saturn 1B with the first crew of three in their CSM was launched on 25 May 1973 on a 28-day mission.

Figure 6 shows the Skylab 2 stack atop the milk stool and mobile launcher on Pad 39B. The Mobile Service Structure, propelled by the crawler, is approaching the Saturn 1B vehicle. Skylab 3 was launched on 28 July on a 59-day flight. The program concluded with the third mission, launched on 16 November and splashdown occurred on 8 February 1974 after an 84-day flight.



Figure 6. The Skylab vehicle on the “milk stool” on Pad 39B on 11 January 1971. The work stands on the Mobile Service Structure are clearly seen as it is propelled to the prelaunch position by the crawler (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Apollo–Soyuz Test Program. KSC then turned its attention to the Apollo–Soyuz Test Program (ASTP) which was a joint program between the Soviet Union and the United States for rendezvous and docking of a Soyuz, which would be launched first, and an Apollo CSM and joint flight of the two docked spacecraft. This called for using another Saturn 1B and Apollo CSM. Liftoff of the Apollo with three astronauts aboard took place on 15 July 1975. The U.S. portion of the successful mission was terminated by splashdown of the Apollo command module on 24 July 1975 (11).

With the launch of ASTP, KSC entered a period of almost 8 years with no launches at Complex 39. Although there were no launches, it was a period of intense activity preparing the center for the Space Shuttle program.

Space Shuttle. The Space Shuttle program, which is described in detail elsewhere, featured a reusable Spacecraft called the Orbiter, boosted into space by three rocket engines in the Orbiter and by two large solid rocket boosters which would be jettisoned after burnout, recovered, and reused. After solid rocket separations, the Orbiter would continue on to orbit using its own engines which consumed liquid oxygen and liquid hydrogen from a huge external tank (ET) which would be jettisoned after engine burnout. The Orbiter would then continue to achieve the desired orbit by using its own onboard thrusters. After the mission was completed, the Orbiter would use thrusters to decelerate from orbital velocity and land like an airplane, although without any engines running.

The mobile launch concept was carried over to the Space Shuttle, so the mobile launcher crawlers, crawlerway VAB, LCC, and the two pads were destined to be used with the new program, even though some required considerable modification. Because the orbiters were to be reusable, a 15,000-foot landing strip, eventually called the Shuttle Landing Facility (SLF), was required. The Orbiters had to undergo considerable refurbishment after return to KSC, so a two-bay building called the Orbiter Processing Facility (OPF) was needed for the refurbishment and for necessary modifications for the next mission, as well as for the initial preparations of each Orbiter as it came on line. Some payloads for Shuttle missions were installed in the OPF. Figure 7 shows the SLF, the VAB, the LCC, the OPF, and part of the crawlerway.

An interesting change of nomenclature, emphasizing the Orbiter's reusability, had occurred with the advent of the Shuttle. Vehicles that were used only once, like the Delta, the Atlas Centaur, and the Titans, were now called "expendable," as opposed to the reusable Shuttle. Heretofore, all launch vehicles, even those used in the manned programs, were in fact expendable. Whether or not the fact that the terms "manned" and "unmanned" were now politically incorrect was involved in the name change is not clear. KSC, however, went to the extent of renaming its Unmanned Launch Operations Directorate the Expendable Launch Vehicles Directorate.

Modifications of the mobile launcher included removing the umbilical tower, reconfiguring the openings in the base to accommodate the different solid and liquid engine configurations, adding two tail service masts to handle connections between the launcher and the orbiter, and adding a hold-down/release mechanism to allow analyzing the liquid engine performance before igniting the solids. A number of modifications were made to the VAB; the most important was the



Figure 7. The Shuttle Landing Facility at the top of this 1983 photograph taken from the southeast. The VAB is prominent in the center, and the Launch Control Center can be seen nearer the bottom of the picture. The Orbiter Processing Facility is visible to the west of the VAB. The Mate/Demate Device is barely visible at the right of the apron between the SLF and the tow way to the OPF. The VAB displays the American flag and the U.S. Bicentennial symbol as a carryover of the Bicentennial “Third Century America” exposition held at KSC in 1976. The Bicentennial symbol has since been replaced by the traditional NASA “meatball” (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

modification of the low bay area to a checkout area for the solid rocket segments and nose cones to prepare them for stacking on the modified mobile launcher in the high bay cells.

The process of stacking started with transporting the mobile launcher into one of the high bay cells with the crawler. The two solid motors were then assembled segment by segment on the launcher. The external tank was then attached to the two solid motors. Last the Orbiter was rolled over to the VAB transfer aisle on its own landing gear, hoisted into position, and attached to the ET. After VAB operations were complete, the crawler returned and transported the entire stack of two SRBs, the ET, and the Orbiter to one of the two pads for final checkout and launch. (See the article on the U.S. Manned Spaceflight: Mercury to the Shuttle elsewhere in this Encyclopedia.) The fact that the SRBs were the basis of the stack, added to the presence of SRB segments in the VAB at almost all times, precluded the use of the office space in the VAB, as contrasted with the Apollo setup where ordnance was installed just before rollout.

At the pads, the modifications included changes in the flame bucket to match the penetrations of the mobile launcher necessitated by the new motor configuration. A fixed service structure (FSS) mounted on the pad surface

replaced the tower on the launchers. Among the features on the FSS are the Crew Access Arm, through which the crews enter and exit the Shuttle—ingress and egress in NASA parlance and the so called “Beanie Cap” which is positioned above the ET and draws off the vented gases from the cryogenic propellant tanks.

Attached to the FSS is the Rotating Service Structure (RSS) that rotates about a large hinge on the FSS allowing it to interface with the Orbiter when required and to rotate to the retired position as launch approached. The RSS provides some protection from the elements for the Orbiter and, more importantly, contains a Payload Changeout Room (PCR) that allows access to the payload bay and allows installing payloads at the pad rather than in the OPF much earlier in the preparation process. Eventually, most users opted to install payloads at the pad, allowing for shorter times in the field for spacecraft crews and less unattended times for the spacecraft.

Modifications were also necessary to facilities away from Complex 39. The O&C building in the industrial area was modified to check out payloads that would be flown repetitively and that would usually be installed in the OPF. These payloads were sometimes called horizontally processed payloads. The most prominent of these was Spacelab, a facility that fits into the Shuttle payload bay and was built in Europe by the European Space Agency (ESA). Although final assembly and checkout of Spacelab before delivery to KSC was accomplished in Bremen in the Federal Republic of Germany, many of the components were manufactured in plants located in other ESA member countries.

An interesting product of the Shuttle era at KSC was the acquisition of a NASA “navy.” As mentioned earlier, the solid rocket boosters were designed to be recovered, refurbished, and reused. After booster separation, parachutes were deployed, and the burnt out rockets descended to the ocean surface, where they would be located and towed back to a new dock behind Hangar AF on the Cape by two specially designed ships, the Liberty and Freedom leased by NASA. They were taken into the hangar, disassembled, and put into condition for rail shipment back to the factory for refurbishment for a subsequent flight.

Shuttle Operations at KSC. The first Orbiter named “Columbia” arrived at the KSC SLF on 24 March 1979. It had ridden piggyback from its California manufacturing plant atop a specially modified 747 aircraft called the Shuttle Carrier Aircraft (SCA). The two vehicles were then towed to another new KSC facility called the Mate/Demate Device (MDD) which was capable of lifting the Orbiter from the attach fittings on the 747 and, after the 747 had been moved away, of lowering it to the surface on its own landing gear. (The MDD can also reverse the process of lifting an Orbiter from the surface to the back of a 747, which has been done when Orbiters have been returned to the factory for major modifications and upgrades.) Columbia was then towed to the OPF to begin launch preparations. The other three original Orbiters were “Challenger,” “Discovery,” and “Atlantis,” all named for ships famous for exploration. An earlier model with no engines was named “Enterprise” and was used for approach and landing tests at Edwards, for vibration tests at MSFC, and as a pathfinder for the new facilities at KSC. After the Challenger was destroyed, an additional Orbiter called “Endeavour” joined the fleet.

Columbia arrived at KSC with considerably more factory style work yet to be accomplished, mostly in the installation of the thermal protection tiles on areas of the Orbiter that would be susceptible to extreme heating during reentry. The original estimate was that about 25% of the approximately 31,000 tiles still needed installation. New tests that raised concern about the capability of many already installed tiles resulted in the removal of thousands of these tiles. This situation severely increased the amount of work to be done at KSC and resulted in substantial delays. Indeed, a second bay of the OPF was turned into a virtual production facility for tiles.

Meanwhile, other components were arriving at KSC. A pathfinder ET had arrived by barge in March. The flight ET arrived on 6 July 1979, and the first of the three Orbiter engines arrived on 10 July. The first solid rocket motor segments arrived in September.

The Orbiter finally was towed from the OPF to the transfer aisle of the VAB on 24 November 1980 after some twenty months in the OPF. In the VAB, it was mated to the ET and the SRBs already on the launch platform. On 29 December 1980, the crawler picked up the stack, and by evening the first Space Shuttle was on its launch pad, ready for preflight operations of the first shuttle mission, called STS-1 (for Space Transportation System). Unlike all other U.S. manned spaceflight programs, whose early launches were accomplished without a human crew, the first launch of Columbia was to be piloted by two astronauts. John Young, veteran of two Gemini flights and two Apollo flights, was the commander, and Robert Crippen, who had no previous space experience, was the pilot. On 12 April 1981, after a 20-second flight readiness firing on 20 February and an aborted countdown on 10 April, the first Space Shuttle lifted off on a mission of 2 days and 6 hours, terminated by a successful landing on a dry lake bed runway at Dryden Flight Research Center at Edwards Air Force Base in California.

A 300-man crew from KSC was stationed at Edwards to service the Orbiter and prepare it for return to KSC aboard the SCA two weeks later. An MDD like that at KSC had been constructed at Edwards for use in the Approach and Landing Test of the Enterprise and to facilitate delivery of the Orbiters from their relatively nearby production facilities in Palmdale (12).

On its first flight Columbia carried no payload per se. It did, however, carry extensive instrumentation called Development Flight Instrumentation (DFI) to measure the Orbiter's environment and performance. The second, third, and fourth flights of Columbia were primarily additional flight tests, but each of the STS-2 and STS-3 missions included a payload for NASA, and the STS-4 mission included a payload for the Air Force.

STS-5, using Columbia, was the first operational mission. It successfully carried into low Earth orbit two commercial communications satellites each with two solid motors, one for achieving a geosynchronous transfer orbit after being separated from the orbiter bay and a smaller one for circularizing the orbit to geostationary. The launch of STS-5 took place on 11 November 1982.

STS-6, the first flight of the second Orbiter, Challenger, carried a larger communications satellite and a different solid motor. It, too, was successful.

On the tenth STS flight, Challenger concluded another commercial spacecraft delivery mission, and the first Orbiter landed at the SLF on 11 February

1984. (The NASA nomenclature system for designating missions was changed starting with this mission, which was called STS 41-B, rather than STS-10.)

Mission 41-D, the twelfth mission, marked the introduction of the third Orbiter, Discovery, into the fleet on a 6-day satellite delivery flight launched on 30 August 1984. The fourth Orbiter, Atlantis, made its first flight in October 1985 carrying a Department of Defense payload as Mission STS 51-J, the twenty-first launch.

By the end of 1985, the STS had successfully accomplished 23 missions with a variety of mission objectives. The launch rate had increased with the addition of new Orbiters, from two in 1981, to three in 1982, to four in 1983, to five in 1984, and to nine in 1985. One successful mission had been flown in early January 1986, when disaster and tragedy struck the program with the destruction of Challenger soon after launch as Mission 51-L on 28 January 1986.

The Challenger loss, and the actions NASA took as a result, caused a significant hiatus in the launch of shuttles. The launch of Discovery on the STS-26 satellite delivery mission occurred more than 32 months later on 29 September 1988. (The nomenclature of missions had reverted to the former system.) By the time flights were resumed, several important policy changes had been made. No shuttle launches would be made at Vandenberg, commercial spacecraft would not be flown on shuttles, and no new DoD missions would be manifested on the Shuttle.

As a result, after resumption of flight, the launch rate did not continue to increase as had been planned, and the mark of nine per year has not been exceeded, even though an additional Orbiter, Endeavour, joined the fleet and made its first flight as STS-49 in September 1992. Parenthetically, the designation as STS-49 does not mean that it was the forty-ninth flight; STS numbers were now assigned with the manifesting of the mission, and the flights did not necessarily occur in that order. For instance, STS-28 was flown after STS-29 and STS-30, all in 1989.

Shuttle Payload Processing. The Shuttle would carry two principal kinds of payloads reusable and attached like Spacelab and deployable, typified by communications satellites. Even that distinction is not absolute because the deployables require cradles, erection devices, and separation systems which were not deployed but remained in the payload bay and had to be removed before the next flight.

In the beginnings of the Shuttle program, the term “cargo” was frequently used to describe the complement of experiments, satellites, attached payloads, and airborne equipment used to elevate and separate deployable payloads carried on a given Shuttle flight. A payload could be either a part of the cargo or, in the case of Spacelab, possibly the entire cargo. Gradually, as the program matured, this distinction between cargo and payload became blurred and though there are still references to cargo, the terms tend to be essentially interchangeable, and the term payload is more common.

Spacelab, a so-called attached payload, was installed in the cargo bay and remained there throughout the mission. It could be flown in either a manned or unmanned configuration. In the manned configuration, it consisted of a pressurized module outfitted with experiments and connected with the crew compartment of the Orbiter by a tunnel. Once the Orbiter and Spacelab were in orbit, the mission and payload specialists could enter the Spacelab through the tunnel and perform the planned experiments; similarly when the flight was nearly complete, the Spacelab crew secured the experiments, left through the tunnel, and went

through deorbit, reentry, and landing in the crew compartment. The unmanned version consisted of a number of pallets containing experiments, which also stayed in the bay during flight, but which did not require hands-on attention. Any operation of the instruments could be accomplished from the aft flight deck of the Orbiter.

Most spacecraft that were intended to be deployed from the Orbiter are processed in some of the same facilities on KSC and on the Cape that had been used for the unmanned programs. One of these facilities, Building AO on Cape Canaveral, contains a large clean room that was used extensively to prepare planetary spacecraft. Most of these payloads were assembled into a cargo in the Vertical Processing Facility, a building that was originally constructed in 1964 as the Pyrotechnic Installation Building for Apollo spacecraft components. It had been previously used with the name Spacecraft Assembly and Encapsulation Facility 1 (SAEF-1) for planetary spacecraft and even housed the sterilization facility mandated to keep the Viking landers from contaminating Mars.

Payloads have historically been processed using two general philosophies. One is the host mode whereby the resident agency, usually the project providing the launch vehicle, provides facilities, services, and coordination functions with the Range and itself to a payload team that will be at the launch site for a relatively short time, sometimes called a launch campaign. The visiting project team will check out the payload for itself, depending on a coordinator and a team from the launch vehicle project, called the Launch Site Support Manager (LSSM) and the Launch Site Support Team (LSST), to see that its needs are fulfilled and, equally important, that it does not run afoul of Range and vehicle project rules and regulations, particularly in the area of safety. This host mode has been used by NASA's Unmanned Launch Operations Directorate (later renamed Expendable Vehicles Directorate) in the conduct of the Delta, Atlas Agena, Atlas Centaur, and Titan Centaur programs. It is appropriate for launch vehicle projects that launch a variety of spacecraft.

An alternate mode is more appropriate for a payload project that has a number of similar spacecraft. In this case, launch site personnel are delegated the task of performing the actual checkout, of course, subject to the requirements of the payload project. This method has been used on the Mercury, Gemini, and Apollo programs.

The Space Shuttle program has used both methods. Repetitive payload programs such as Spacelab were handled by resident NASA personnel. Generally deployable satellites that include a number of significantly different configurations are treated in the host mode. LSSMs were generally assigned in either case.

Payload processing was certainly not a new thing at either the Cape or at KSC, but the advent of the STS, its reusable delivery method, and the concept of reusable payloads brought about a new look to the process (13,14).

The International Space Station and KSC. The construction of the International Space Station (ISS) placed new requirements on KSC and resulted in the erection of a new facility at KSC, the Space Station Processing Facility (SSPF). The building contains processing bays, an airlock, control rooms, laboratories, logistics areas, office space, and a cafeteria. After the construction of the SSPF had begun, a decision was made to perform leak checks on pressurized modules of the International Space Station at KSC. One of the altitude

chambers in the O&C building which had been used in the Apollo, Skylab, and ASTP programs was refurbished and turned over to ISS operators in 1999 (15). These altitude chambers were deactivated after the ASTP mission in 1975, but they remained in the building unused because there was no requirement for them for Space Shuttle payloads and because removing them would have been very expensive and disruptive.

The new SSPF has been used for the prelaunch checkout of segments of the ISS and will continue to support the program with checkout of other segments of the ISS, experiments, and resupply items that will be transported on the Shuttle throughout the life of the ISS.

Accommodating the Public

NASA has been aware of the requirement in its charter to keep the public informed of its activities. From its beginning, it has given the press information about and access to launches and other major events. In addition, KSC and other centers operate Visitors Centers, where the public is allowed to view exhibits and actual space hardware. These are normally located outside the gates in areas for which no credentials or passes are required. One of the features at the KSC Visitor Center is the Rocket Garden, which includes full size replicas (in some cases actual surplus flight hardware) of space launch vehicles of the past. A full size replica of a shuttle orbiter can be toured, and full scale exhibits of the external tank and the solid rocket boosters can be viewed. Also at KSC, there is the additional feature of bus tours throughout the center that allow the public, for a fee, to ride through the center and to some NASA sites on Cape Canaveral with the option of stopping at certain sites, such as the SSPF observation gallery, an observation tower from which one can see the entire Complex 39, and a Saturn V Apollo building devoted to Apollo, Skylab and ASTP. Within this building is actual Saturn V and Apollo hardware displayed horizontally with the stages and spacecraft elements separated. The KSC Visitor Center, which is operated by a concessionaire, consistently ranks among the top tourist attractions in Florida.

The Air Force, which faces a different set of security concerns, has also accommodated the public with its Air Force Space and Missile Museum. The museum is located within the boundaries of the Cape at deactivated Complex 26, the complex from which America's first Earth satellite was launched. The blockhouse is used for a display area, and many missiles and space vehicles are erected on the grounds in a Rocket Garden similar to its namesake at the KSC Visitor Center. It includes early ballistic and aerodynamic missiles in addition to some space launch vehicles. Admission is free, but operational and safety concerns limit visiting hours. Adjacent to the Air Force Museum is the deactivated Complex 5/6, the site of the Mercury Redstone launches. A Mercury Redstone and a Jupiter C are displayed in their gantries and can be viewed on drive-through tours of this facility.

BIBLIOGRAPHY

1. Cleary, M.C. *The 45th Space Wing: Its Heritage, History & Honors 1959–1996*. 45th Space Wing History Office, Patrick Air Force Base, FL, 1997.

2. Harris, G. The Year the Rockets Came. *Smithsonian Institution* Washington, DC, April/May 1999.
3. KSC Public Affairs. The Kennedy Space Center Story. Kennedy Space Center, FL, 1991.
4. KSC Public Affairs. The Kennedy Space Center Story. Kennedy Space Center, FL, 1991.
5. KSC Public Affairs. *Spaceport News* Kennedy Space Center, FL, May 28, 1999.
6. Cleary, M.C. The 45th Space Wing: Its Heritage, History & Honors 1950–1996. 45th Space Wing History Office, Patrick Air Force Base, FL, 1997.
7. Cleary, M.C. The 6555th: Missile and Space Launches Through 1970. 45th Space Wing History Office, Patrick Air Force Base, FL, 1991.
8. Cleary, M.C. The 45th Space Wing: Its Heritage, History and Honors 1950–1996. 45th Space Wing History Office, Patrick Air Force Base, FL, 1997.
9. Office of Commercial Space Transportation. Hazard Analysis of Commercial Space Transportation. U.S. Department of Transportation, Washington, DC, 1988.
10. Benson, C.D., and W.B. Faherty. Moonport: A History of Apollo Launch Facilities and Operations. The NASA History Series, Washington, DC, 1978.
11. KSC Public Affairs. The Kennedy Space Center Story. Kennedy Space Center, FL, 1991.
12. KSC Public Affairs. The Kennedy Space Center Story. Kennedy Space Center, FL, 1991.
13. Neilon, J.J. Preflight and Postflight Processing of Spacelab Missions at KSC. *33rd Cong. Int. Astronaut. Fed.*, Paris, France, 1982.
14. Neilon, J.J. Processing Cargoes for the First Two Operational STS Flights at KSC. *34th Cong. Int. Astronaut. Fed.*, Budapest, Hungary, 1983.
15. KSC Public Affairs. *Spaceport News* Kennedy Space Center, FL, March 5, 1999.

READING LIST

- Brooks, C.C., J.M. Grimwood, and L.S. Swenson, Jr. Chariots for Apollo: A History of Manned Lunar Spacecraft. NASA History Series, Washington, DC, 1979.
- Cleary, M.C. The Cape Military Space Operations 1971–1992. 45th Space Wing History Office, Patrick Air Force Base, FL, 1994.
- Hacker, C., and J.M. Grimwood. On the Shoulders of Titans: A History of Project Gemini. NASA History Series, Washington, DC, 1977.
- Swenson, L.S. Jr., J.M. Grimwood, and C.C. Alexander. This New Ocean: A History of Project Mercury. NASA History Series, Washington, DC, 1966.

JOHN J. NEILON
Cocoa Beach, Florida

ELECTROMAGNETIC PROPULSION

Introduction

Sporadic attempts have been made for more than 150 years to use electrical energy to launch projectiles at high velocities. Throughout the world during the

last 20 years, most of the funding invested in electromagnetic (EM) launcher development has been for military applications. In most cases, the perceived advantage has been attaining higher velocities than can be achieved with propellant guns, leading to enhanced lethality. The elimination of gun propellants would also reduce vulnerability as well as allowing an increased number of stored rounds and enabling extended missions. The EM rail gun has been the preferred choice for this application because it has already repeatedly demonstrated muzzle velocities well above those of conventional guns—typically 2500 m/s versus 1750 m/s. The military challenge is to fit the electric pulsed power equipment into a battleworthy vehicle, especially if it needs to be compact. The main potential applications are direct fire tank guns, long-range artillery, and air and missile defense.

Several electrical or electromagnetic (EM) concepts have been suggested, as described later, but their potential for space applications is, as yet, limited. However, the technological bases are now being created on which future EM launch systems for space exploration could be based in the twenty-first century. Such systems could include launch to space from the surface of Earth, the Moon or asteroids, Maglev launch assist for large rocket systems and, possibly, EM thrusters for space transportation. Electrical ion engines have been used in space, and the Hall plasma thruster is under development for NASA.

The interest in electric launchers for military applications has led to improved understanding and advances in many technical areas. Most military applications for EM launchers require gigawatt power levels for pulse lengths of a few milliseconds to accelerate a useful projectile in a launcher of acceptable length. Coupled with the need for compact and robust equipment, this is a challenging goal for the pulsed power system. The desire to achieve muzzle velocities significantly greater than those of conventional guns—from 2–3 km/s—also requires developing the launcher and projectile. Recent progress in these areas could be the basis for future space applications.

History

There are isolated reports of efforts to develop EM propulsion technology in the nineteenth century, but the major developments were in the twentieth century. Pioneering efforts were undertaken in Norway in 1901–1903, in France during World War I, in the United States in the 1930s, and in Germany during World War II. In the second half of the twentieth century there were notable contributions from many countries, including Australia, Britain, China, France, Germany, Korea, Russia, the Ukraine, and especially the United States.

The first known scientific account of an EM launcher was provided by Prof. Page of the Colombian College, Washington, D.C., in 1845, only 10 years after Faraday's pioneering studies on electricity (1). Page's description of the electric launcher stated,

Another curious instrument is the galvanic or magnetic gun. Four or more (helical coils) arranged successively, constitute the barrel of the gun, which is mounted with a stock and breech. The (iron) bar slides freely through the (coils) and, by means of a wire attached to the end towards the breech of the gun, it makes and breaks the

connection with several coils in succession and acquires such a velocity as to be projected to the distance of forty or fifty feet.

The well-known Norwegian inventor, Birkeland, developed an EM launcher using multiple coils (2). He obtained the first patent for an electric gun, formed a company to exploit the technology, and successfully tested two 4-m long induction launchers in 1902. However, his widely advertised public demonstration in 1903 spectacularly failed to live up to the claims that the projectile would be launched without flash or noise, and within weeks, he moved on, in this case successfully, to a different technical career.

Seeking a different approach, Fauchon-Villeplee started studies on EM rail launchers in France during World War I (3). In 1917, he successfully constructed a small working model powered by batteries. The principles of his direct current (dc) electric rail gun were different from those of the coil, or solenoid, launchers. A cross section of the basic configuration taken from one of his patents is shown in Fig. 1. The structure of the gun was iron, and it had two pole pieces to concentrate the magnetic flux into the region where the projectile was located. The magnetic field was energized by copper windings that extended along the length of the barrel and used the same current as fed into the fins of the projectile through sliding contacts. The lightweight projectile was a plastic tube that had a wooden tail in which copper fins transferred current via a sliding contact with the rails. Fauchon-Villaplee proposed that an EM cannon could be built to launch a mass of 100 kg at 1600 m/s. This formidable gun, shown mounted on a railway bogie in Fig. 2, was supposed to have a muzzle energy of 128 MJ—an energy level that has still not been achieved. When World War I ended, these projects were abandoned, but the work did not go unnoticed. A 1923 report by Korol'kov of Russia (4) critiqued the physics and engineering calculations of Fauchon-Villeplee and commented positively and negatively on the claimed advantages of electric guns:

- Electric guns fire without smoke and with virtually no noise.
- Electric gun efficiencies are considerably higher than those of conventional guns.
- Electrical guns use fuels whose energy is greater than explosives and are less expensive.

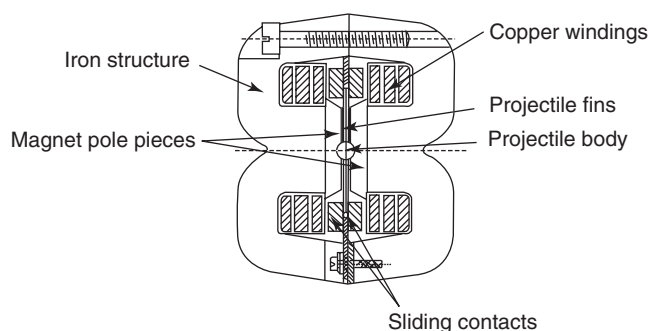


Figure 1. Cross section of Fauchon-Villaplee's iron-cored rail gun barrel (ca. 1918).

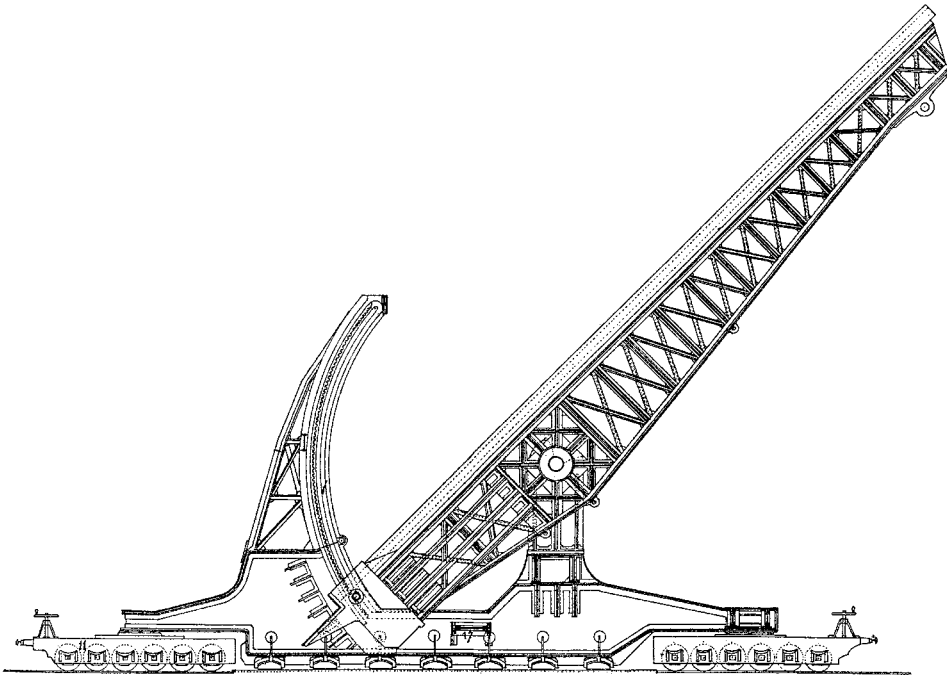


Figure 2. Fauchon-Villaplee's long-range gun concept (ca. 1918).

- The force propelling the projectile is applied along the entire projectile, not just its base.
- The velocity and range of the projectile can be changed by varying the voltage without changing the angle of elevation of the barrel.
- There is no (powder) charge in electric guns, and there cannot be a premature explosion in the gun barrel.
- The power of electric guns...depends only on the instantaneous energy of the electrical source.

MacLaren's U.S. patent of 1921 described a traveling magnetic wave induction launcher powered by a high-frequency, three-phase generator (5). In this, the traveling magnetic wave induced currents in the projectile and provided the EM force that accelerated the projectile through the barrel. In the 1930s, Professor Northrop undertook experiments at Princeton University with similar launchers powered by motor-generators and capacitors. The work was described in a novel written by Northrop¹ (6) in which the theme was a journey to the Moon using an EM coil launcher in place of the conventional cannon proposed by Jules Verne. Northrop provided considerable detail on the engineering calculations, as well as several photographs illustrating the equipment needed—some of which were retouched to appear more impressive.

¹Using the pen name "Pseudoman."

During World War II, a group led by Hansler in Germany was involved in developing EM launchers (7). The experiments started with induction launchers but because little success was achieved, refocused on rail guns. Hansler's first rail launcher tests in September 1944 used an iron barrel similar to Fauchon-Villaplee's that was energized by windings to create a transverse magnetic field to give the $J \times B$ forces needed for accelerating a projectile (Fig. 3). A velocity of 1080 m/s was achieved using a spring-loaded arrangement to ensure good electrical contact between the projectile fins and the rails. To meet the wartime needs of defending against high-altitude bombing raids, the development of a 40-mm EM gun with a high velocity (2000 m/s) and rate of fire of 72 rounds/min. was proposed. The initial design had a 10-meter iron barrel with a 3-tesla magnetic field. However, Hänsler soon realized that the performance required could be achieved only with an air-cored barrel that had a much higher magnetic field, as shown in Fig. 4. This was designed, but not built, when the war ended. Modern EM rail-gun launchers are of this type, and operating currents are so high that pulsed magnetic fields > 20 T are generated, making the presence of iron irrelevant. Starting in May 1946, the Armour Research Foundation in the United States evaluated Hansler's work. A comprehensive document was created based on reports, blueprints, and equipment from Germany (8). After evaluating the material and performing calculations, the report concluded that the anti-aircraft EM rail gun was technically feasible.

Relatively few efforts were undertaken in the aftermath of World War II, with one or two exceptions (9,10), until rail-gun experiments were undertaken at the Australian National University in the early 1970s based on the concept of macroparticle acceleration to high velocities for meteorite impact studies (11). These experiments led to a growth of interest in the United States and also caused non-U.S. efforts to accelerate. Most of these efforts were for defense applications, but launch-to-space was studied by MIT and NASA.

The early pioneers of electric gun research would be impressed to see what has been accomplished in the last decade. Pulsed power supplies using capacitors and rotating generators have provided currents in excess of 3 MA to EM rail launchers. Velocities of more than 6 km/s have been achieved in rail guns using

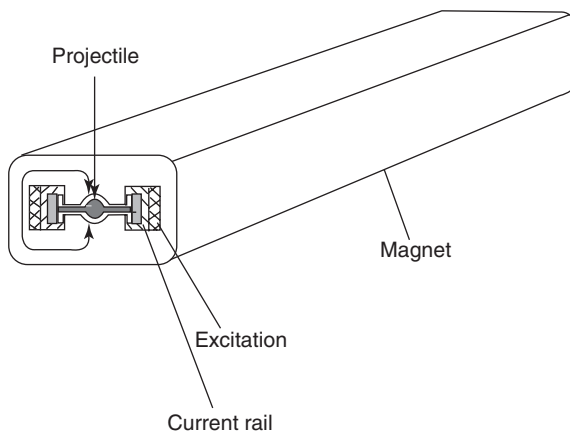


Figure 3. Sketch of Hansler's iron-cored rail-gun (ca. 1944).

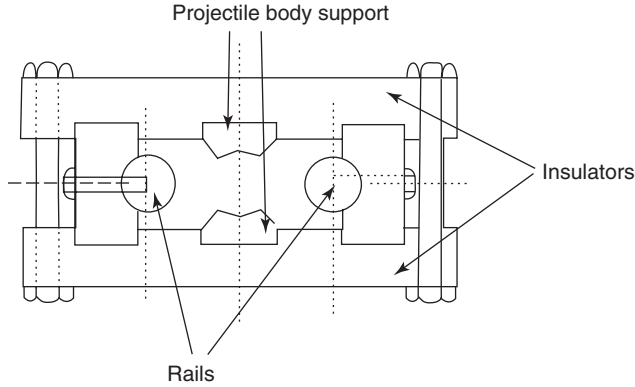


Figure 4. Hansler's air-cored rail-gun cross section.

projectiles of a few grams, and muzzle energies of 5–8 MJ have been achieved for larger masses at velocities up to ~ 4 km/s. Induction launcher technology has been demonstrated for low velocities (some 100's m/s), but the highest confirmed velocity is 1 km/s.

Electric Launcher Configurations

Rail Launcher Parameters. Rail launchers and coil launchers rely solely on electromagnetic (EM) forces to accelerate projectiles to high velocities. Most single-stage laboratory EM rail launchers operate at efficiencies of 30% or less, so that substantial energy is needed at the breech of the launcher for each shot. Modifications to improve launcher efficiency are described later. Some of these are more appropriate for launch-to-space than for ordnance applications.

Most experimental and theoretical studies of EM rail launchers during the last 20 years have used the simple air-cored rail-gun arrangement proposed by Hansler. In this, a short pulse of high-current electrical power is fed to the breech of a two-rail configuration in which the projectile is accelerated by the Lorentz ($J \times B$) force (Fig. 5). This configuration has operated successfully across a wide range of conditions.

Because the induced magnetic field in the air-cored launcher bore is created by the current I passing through the rails, the $J \times B$ accelerating force (F_{EM}) can be expressed as

$$F_{EM} = 0.5L'I^2, \quad (1)$$

where L' is the inductance gradient ($=dL/dx$) of the rails in the vicinity of the projectile. L' is a relatively invariant quantity that depends on rail-gun bore geometry, more accurately, the relative fraction of the bore cross section occupied by the rails: a typical value is $\sim 0.5 \mu\text{H/m}$. L' is small, so a large current is required to accelerate a significant mass m to a high velocity because

$$F_{EM} = 0.5L'I^2 = ma. \quad (2)$$

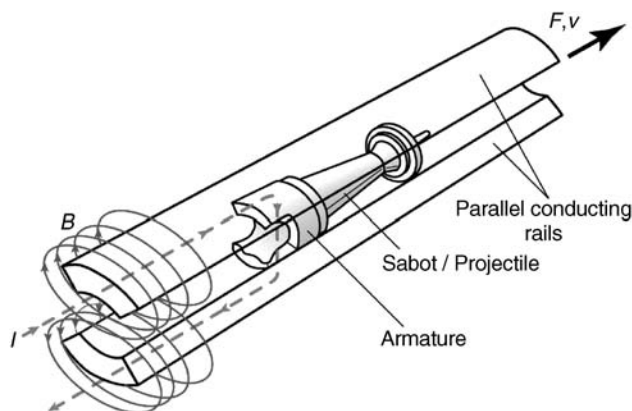


Figure 5. The two-rail, breech-fed rail-gun arrangement. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Currents of several megamperes have been successfully used in 90-mm bore rail guns in laboratory experiments.

Because the accelerating force depends on I^2 , it is independent of current direction and, in principle, an alternating current (ac) could be used. Some early systems were of this type, but shorter barrels and a better acceleration profile can be achieved by using a unipolar (dc) current pulse. Modern rail launchers use a current pulse that has a fast rise to a constant current for the majority of the acceleration, followed by a decrease in current near the muzzle exit to minimize muzzle flash. This “flattopped” pulse shape can be achieved using several short pulses in sequence from each section of a subdivided energy store or generator. Using this approach, the subpulses must be correctly timed and overlapped to achieve the desired total pulse shape (12).

One major source of inefficiency in the simple rail launcher is the magnetic energy remaining in the barrel at the instant of projectile launch. For an idealized constant current pulse, the inductive energy left in the barrel at launch equals the projectile muzzle energy so that the launcher efficiency², even without other losses, cannot exceed 50%. Reducing the current at the muzzle exit decreases the magnetic energy stored in the barrel but increases the barrel length. The need for compact power supplies has led to the invention of alternative launcher configurations that may have higher efficiency, as described later. None of these has yet been widely studied, and further research is required to determine whether the advantages of these concepts outweigh their disadvantages.

Numerical models are used for the detailed design of EM rail launchers, but simple expressions can illustrate the fundamental principles and provide approximate engineering parameters, as shown here.

When the fundamental definitions of acceleration a and velocity v are integrated for the condition of constant acceleration \bar{a} , they yield the well-known result that: “*work done (or energy) equals force times distance.*” For an EM

²The efficiency is the kinetic energy of the projectile at the muzzle exit divided by the electrical energy supplied to the breech of the launcher.

launcher this can be written as

$$E_{\text{muz}} = 0.5mv_{\text{muz}}^2 = \overline{F_{\text{EM}}}s, \quad (3)$$

in which s is the launcher barrel length and \overline{F} is the average EM force. Assuming that the force is constant across the cross-sectional area A of the barrel, the accelerating force is also

$$F_{\text{EM}} = pA. \quad (4)$$

In EM launchers, as in propellant guns, the accelerating pressure p does not remain constant during the launch process, unless special design approaches are taken. The measure of how close the acceleration is to the ideal is the piezometric efficiency, η_P

$$\eta_P = p_{\text{max}}/p_{\text{ave}}. \quad (5)$$

For a propellant gun, $\eta_P > 2$, but simple rail launchers have operated with $\eta_P \sim 1.5$, and operation to 1.3 or lower may be possible. Even lower values may be feasible with the “distributed” EM launchers described later.

From Equations 3–5,

$$E_{\text{muz}} = p_{\text{max}}As/\eta_P, \quad (6)$$

showing that, for a given barrel length and area, the muzzle energy is determined only by the average pressure \bar{p} , or by the maximum pressure, p_{max} and the piezometric ratio.

Using $I^* = I/h$ as a measure of the linear current density on the inner surface of the rails that carry current (where h is the height of the rail at the inner bore from one insulator to the other), Equation 2 can be written as

$$F_{\text{EM}} = 0.5L'I^{*2}h^2. \quad (7)$$

Combining Equations 4 and 7 and using $A = h^2$ for a square barrel yields

$$p_{\text{sq}} = 0.5L'I^{*2}. \quad (8)$$

Typical values in this expression are $L' = 0.5 \mu\text{H/m}$ and, for a copper rail, $I^* \sim 40 \text{ kA/mm}$, to give $(p_{\text{sq}})_{\text{max}} = 400 \text{ MPa}$. This is the highest pressure that a simple EM rail launcher can achieve. Substituting $(p_{\text{sq}})_{\text{max}}$ in Equation 6 allows rail launcher parameters to be estimated.

For simplicity, the above analysis assumed a square rail launcher bore. Other geometries (e.g., rectangular or circular) have been investigated experimentally and analytically. Calculating L' is complicated in complex barrel cross-sections and requires knowledge of the variation of current density and temperature around the periphery of the rail (13).

The maximum acceleration pressure achievable in a simple rail launcher is lower than the peak values achieved in powder guns, but useful performance levels can be still achieved with a rail launcher, even within these pressure constraints, as shown in Table 1.

Table 1. Maximum Launch Energy (MJ) Achievable in a Simple Square Bore Rail Launcher

Barrel length, m	Bore height, mm			
	80	100	120	140
4	7.9	12.3	17.7	24.1
5	9.8	15.4	22.2	30.2
6	11.8	18.5	26.6	36.2
8	15.8	24.6	35.4	48.2
10	19.7	30.8	44.3	60.3

Some techniques for increasing the achievable pressure have been sought, as described later, but they are still limited by Equation 8. Increasing I^* requires different rail materials, but the potential benefits of using high-temperature materials is largely offset by their poorer electrical and thermal properties, as can be seen from Table 2.

For example, although tungsten, has a melting point three times higher than copper, it will reach surface melting when $I^* = 56 \text{ kA/mm}$, which is only a 37% improvement over copper. In the absence of fundamental material improvements, it may be necessary to use other rail launcher configurations that have higher L' values, as described later.

Augmented Rail Launcher. In early concepts, external magnets—either superconducting (14) or permanent (15)—were proposed to provide a steady magnetic field that would permeate the entire launcher bore volume to add to the induced magnetic field produced by the applied current. However, such fields are generally small compared with the pulsed field resulting from the main current in large bore launchers, and this technique has now been replaced by additional rails close to the main launcher rails to enhance the magnetic field in the launcher bore. The augmenting rails may be energized either by (1) using all of the current that goes through the main rails, (2) using a separate supply, or (3) taking part of the main current and diverting it through the augmenting rails. The first option has generally been used experimentally because it is simple and only one power supply is needed.

Table 2. Limiting Rail Currents for Several Materials

Rail material	I_{max}^* , MA/m
Copper	40.9
Aluminum	25.3
Silver	32.8
Tungsten	56.2
Graphite	35.0
Stainless steel	32.7
Molybdenum	45.5
Elkonite (67% W, 33% Cu)	65.5

The advantage of augmentation is that the current in the main rails and armature is reduced for a given accelerating force, thereby decreasing bore erosion and easing the task of the armature designer. In addition, the power supply can operate at lower current³. For single augmentation (one augmenting turn in addition to the rails in the barrel bore), $L' \sim 1 \mu\text{H/m}$; double augmentation yields $L' \sim 1.2$ to $1.3 \mu\text{H/m}$. Singly and doubly augmented barrels have been successfully demonstrated with small-bore rail guns (16,17). Augmentation has not been widely considered for military applications because the launcher is larger and heavier than a simple barrel, complicated breech and muzzle current connections are needed, and additional magnetic energy is stored inductively in the barrel at the muzzle exit. However, this may not be a concern for a fixed launcher, as for a launch into space.

DES Rail Launcher. A second alternative to the simple rail gun is the distributed energy store (DES) launcher. In this, multiple power-feed points are located along the rails so that energy can be fed from several independent energy sources (18,19). This allows the maximum current to the projectile armature to be provided at just the right time in the acceleration process to maximize the EM force on the launch package.

To supply this launcher, the energy source⁴ needs to be segmented into separate sections that are energized sequentially as the projectile passes. The magnetic energy is switched into each stage of the launcher exactly when the projectile reaches that stage and is switched off after the projectile leaves. Therefore, the amount of inductive energy left in the barrel when the projectile leaves the muzzle is only that from the last one or two sections. If there are n sections, only $\sim 1/n^{\text{th}}$ of the usual amount of inductive energy remains in the barrel, so that the overall launch efficiency is much improved. This concept may be especially advantageous for a long static system, such as for launch to space. For a military system in which it is necessary to slew the barrel rapidly, the need for multiple power leads to transfer current to each stage from many independent power supplies is a significant disadvantage.

Multirail Launcher. The third alternative version of the simple rail gun is the multiple rail arrangement (20). In this concept, the rails of the simple rail gun are subdivided, so that one- n^{th} of the normal current passes through each rail. The rails are electrically insulated from each other and are electrically in series so that a low-current, high-voltage source can be used. The current required from the power supply is divided by the number of rails, but there is a corresponding increase in the applied voltage. Electrical shorting prevents this technique from being applied with plasma armatures, but it can be used with solid armatures if arcing contact between adjacent rails can be prevented. The concept has been demonstrated experimentally (21,22) but has received relatively little attention because of the concern about the ability to *always* prevent arcing. If arcing occurs between the rails or within the armature, the power supply could be electrically shorted out with disastrous consequences, unless precautions were taken.

Staged Transiently Augmented Rail (STAR) Launcher. The STAR concept (23) combines the features of an augmenting magnetic field and the DES

³Although at the expense of an increased voltage.

⁴And, in some configurations, the barrel rails.

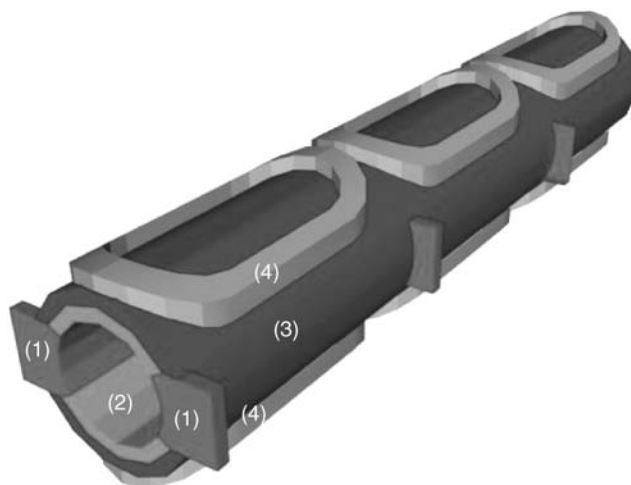


Figure 6. STAR—staged transiently augmented rail gun—module, showing the rails (1), insulator (2), containment (3), and pulsed augmenting magnet coils (4).

power feed. The augmenting field is provided by a series of independent coils that are energized only when the projectile is nearby. Figure 6 illustrates a three-section version of such a launcher; many such sections would be arranged in series for an Earth-to-space launcher. The energy used in these coils (less losses) is then transferred from coil-to-coil as the projectile accelerates along the barrel. The energy required for augmentation is therefore much less than in the usual augmentation concept. This arrangement allows the main rail current to be reduced, which is beneficial for the power supply and the barrel. The total launcher efficiency will also be better than that of the simple rail launcher because the augmentation is transferred forward with the projectile, rather than permeating the entire barrel structure. Although STAR is merely a concept at present, high-field pulsed magnets have been tested at magnetic fields up to 68 tesla (24)—a level that is significantly higher than that used in present EM launchers. The STAR concept is most appropriate for a long static system and is a possible candidate for launch to space.

EM Induction Launchers. In principle, having no direct sliding contacts and operating with relatively low currents, EM induction launchers promise high efficiencies and contactless launch. Operating velocities of a few 100 m/s have been achieved by several groups, but there have only been two reports of velocities near or above 1000 m/s (25,26). As a consequence, there has been little funding for research on high-velocity launch applications, and this technology lags that of rail launchers. However, the approach may be useful for lower velocities and could replace the first stage of a rocket system.

Several induction launcher approaches have been studied (27). The barrel-type launcher consists of two parts—a stationary array of drive coils and a moving conductive projectile. The stationary coils generate a rapidly varying magnetic field that induces currents in the projectile. The interaction between the applied magnetic field and the induced magnetic field creates an axial propelling force and, in some configurations, a radial centering force that eliminates

the need for a contacting guide rail system. The geometry can be coaxial or flat channel, and the induction can be pulsed or traveling-wave. The flat arrangement could be applied to the wings of a projectile in a toroidal configuration.

The single-stage pulsed induction launcher consists of a coaxial, single, fixed, drive coil and movable conduction ring. When the drive coil is pulsed with current i_1 , an induced current i_2 is created in the conduction ring that generates a magnetic field in the direction opposite to that in the drive coil. The mutual repulsion between these two coils causes acceleration of the ring. The EM accelerating force is

$$F = i_1 i_2 \frac{dM}{dx}, \quad (9)$$

where M is the mutual inductance and x the distance between the coils. High velocities and extremely high accelerations were reported with close-coupled, single-stage repulsion launchers of this type (28). A modified arrangement uses a smaller movable ring that can be nested inside the drive coil. This can be further modified by adding additional drive coils to create the barrel of a launcher that can be operated either synchronously or asynchronously. In the synchronous mode, only one drive coil is energized at a time and is synchronized by a feedback switching control system to match the movement of the accelerated coil or cylinder. The pulsed magnetic field from the driving coils induces an azimuthal magnetic field in the rear of the inner coil or cylinder in a direction that opposes the applied field, so that the inner coil is accelerated along the launcher. To maintain a unidirectional azimuthally induced current in the projectile, the current in all of the drive coils must be in the same direction. In strictly synchronous operation, the dc component of the current induced in the projectile decays with time, so that the direction of induced current flow has to be reversed.

Another system is the traveling wave induction launcher in which alternating currents are separately fed to a number of the drive coils across a region that is longer than the projectile being accelerated. This creates a traveling magnetic field that travels with a velocity v :

$$v = 2\tau f, \quad (10)$$

where 2τ is the wavelength of the magnetic wave and f is the frequency of the current in the drive coils. The traveling wave induces azimuthal currents in the projectile and an associated magnetic field pattern that is similar to that generated by the drive coils. However, the induced magnetic field pattern is displaced relative to the driving field pattern, and the projectile is "dragged" along the barrel at a velocity close to that given by Equation 10. The advantages of this arrangement are that the propelling force acts more uniformly across the entire body of the projectile, a radial centering force acts to keep the projectile centered in the barrel, and the need for exact synchronization between the projectile and the drive coils is eliminated. Disadvantages include the need to divide the barrel into several sections that are fed with different frequencies so that the slip velocity and thermal losses can be minimized.

Energy Storage and Pulsed Power

Energy Requirements and Options. From the earliest days of electric launcher research, it has been understood that supplying the required high-power pulses of electricity is a critical issue. Conventional guns and rockets have reached high level of capability after hundreds of years of development in which very energetic propellants have been created. Energy densities in propellants are impressively high—several MJ/kg. Achieving similar energy densities in a system that stores and transfers electrical energy is not yet possible, as can be seen from Table 3. Some of the other benefits of electrical systems such as higher velocities and improved safety due to the absence of propellants offset this disadvantage. In many situations, the capability to easily control the electric launcher to provide the required muzzle energy offers a benefit that is not available from conventional propellants.

The energy needed for an EM launcher is determined by the mission requirements. A notional set of input energy requirements for several launcher technologies for a muzzle energy of 20 MJ is given in Table 4. Other applications, such as an EM catapult for aircraft launch or launch to space, may require much more energy.

It is necessary to provide *average* power to the electrical system at a rate that matches the launch rate. The prime mover⁵ is typically coupled to a generator and operates through a transformer/rectifier to provide electrical power to the energy storage system. The availability of stored energy in a flywheel or battery pack can permit a few launches to occur quickly without energizing the main prime mover power system for certain applications. The design of this subsystem depends on the type of launcher and the firing rates required. For high firing rates or continuous operation, it is necessary to cool the components.

Capacitor Banks. Most early electric launcher experiments used capacitor banks because they are inexpensive in small sizes and the components are widely

Table 3. Electrical Energy Storage Options

Device	Storage mechanism	Defining equation	Assumption	Limiting values	Energy density, MJ/m ³
Capacitor	Electrostatic	$E = \frac{CV^2}{2}$	High-energy-density plastic film	$E_{op} = 400$ $V/\mu m \quad \epsilon_r = 10$	7
Rotor	Inertial	$E = \frac{I_m \omega^2}{2}$	High-speed composite/conductor rotor	$\rho = 1500$ kg/m^3 $v_{periph.} = 600$ m/s	135
Inductor	Magnetic	$E = \frac{B^2}{2\mu_r \mu_0}$	High-field air-cored inductor	$\mu_r = 1 \quad B = 40$ T	640
Battery	Electro-chemical		Li ion	3.5 volts	4000
Propellant or explosive	Chemical		High-energy-density materials	Few eV per bond	5000–10,000
Diesel fuel	Chemical		“Free” oxygen from atmosphere	More exothermic bonds	40,000

Table 4. Input Energies Needed for 20-MJ Muzzle Energy

Electric launcher type	Energy input per shot, MJ	Peak current, kA	Peak voltage, kV	Pulse length, ms	Minimum average instantaneous power, GW
EM railgun	40	4000	10	6	7
ETC gun	4	150	16	4	1
ETC igniter	0.4	50	10	1	0.4

available. Early experiments used banks with an oscillating output current whose characteristic time period was matched to the transit time of the projectile through the rail gun. Because this results in substantial current variation and very nonuniform acceleration, it is generally not preferred. In some cases, a high-frequency, rapidly ringing bank was used, and the averaged fluctuating current was accepted as an inexpensive way to undertake experiments. A better arrangement is to “crowbar” the circuit after the current has reached its peak value, which extends the high-current portion of the output pulse. To achieve this, a closing switch connects the current to the gun, and a diode is used in the crowbar leg with another closing switch.

Even when a single capacitor bank is “crowbarred,” current droop can cause inefficient energy transfer for a launch time that is longer than the characteristic decay time of the bank. The solution is to subdivide the capacitor bank into separate modules that can be independently triggered but are connected together at the current feed point into the launcher breech.⁶ A practical benefit of the modular approach is that repair and maintenance can be carried out on some modules while the remaining units continue to operate. The design of individual modules and the way in which they are connected depends on the type of gun and the duty required. A flat-topped current pulse can be created for a rail gun by triggering several modules sequentially to match the transit time of the projectile through the barrel. To get the desired high initial acceleration, a high-energy module is fired initially, followed by subsequent discharge of smaller modules. Because the projectile is traveling at higher velocity further down the barrel, later modules have smaller inductors in series with the capacitor modules to achieve faster rise times (12).

To ensure high efficiency, the output switches that connect the modules to a single load connection (the launcher breech) must prevent current from later modules from being routed back into modules that discharged earlier (29,30). If this is not done, considerable stored energy may be lost resistively, as it transfers from one module to another without ever reaching the launcher. Capacitor voltages are generally in the range of 10–15 kV, but values of 22, 44 kV, and higher have been used for some experiments. Such values necessitate close attention to insulation integrity and component locations to prevent electrical breakdown.

⁵Usually this is a gas turbine or internal combustion engine.

⁶For a single point breech-fed simple rail gun.

Table 5. Plastic Film Dielectrics

Material	Dielectric constant	Breakdown voltage, V/ μ m	Stored energy Density, MJ/cum
Polypropylene	2.1	200	0.37
Polyethylene	3.25	300	1.3
Polyvinylidene fluoride	10	400	7
Siloxane	9.3	~ 400	$\sim 2.5^a$

^aMeasured with composite materials that had siloxane on a kraft paper backing for strength.

The capacitors in which energy is stored are the single largest contributor to the size and mass of the total system. The energy stored in a dielectric is proportional to the dielectric constant ϵ_r and to the square of the operating electric field strength E_{op}^2 , which is chosen as near as possible to the breakdown strength of the dielectric material while ensuring a low probability of spurious breakdown. Capacitors that use paper dielectric with foil electrodes were standard for many years. Plastic films such as polyethylene (PE), polypropylene (PP), and polyvinylidene fluoride (PVDF) provide better performance (see Table 5).

The performance of PE and PP capacitors is well characterized, and capacitors made from these materials are widely available. Units made from PVDF have been developed in the last decade. Although offering high energy density, this nonlinear material is more difficult to characterize, less efficient in operation, and more difficult to integrate into a complete system (31). As a result of these disadvantages, there is much to be gained if new dielectric materials could either be discovered or created. Recent small-scale experiments have been performed using siloxane-based materials with promising results (32). However, substantial investments are needed to bring a new capacitor material to reliable large-scale manufacture, and few sponsors have yet been found.

It is important to consider the entire capacitor system when component development is planned. If extremely high-energy-density dielectric materials were to become available, the volume of the system would then be dominated by other components, such as switches or inductors. Therefore, these components also need to be developed to reduce their size and weight. Integrated systems, in which component functions are combined, rather than by being individually optimized, may have significant benefits (33).

Rotating Machines—Homopolar Generators. Laboratory studies of electric launchers were undertaken in the 1970s and 1980s using homopolar generators (HPGs) which operated as electrified flywheels to deliver a low-voltage but high-current output (34). The HPG voltage, typically 100–200 V for an iron-cored machine, is much less than that required to drive an electric gun directly. Therefore, it is necessary to add a pulse compression stage in which energy is transferred into an inductor where it is temporarily stored as magnetic energy before delivery to the launcher at a higher voltage. Getting the energy into the inductor is relatively simple, but getting the energy out of the inductor to the gun can be difficult. An opening switch is needed that can carry a current of megamperes during inductor charging, but which can open quickly when transfer of

the current into the gun is required. Such switches are not readily available, although laboratory versions have been developed. The need for the opening switch and storage inductor has caused the HPG to go out of fashion during the last few years in favor of other rotating machines.

Pulsed Alternators. Synchronous ac generators of very large sizes (>1400 MW) operate successfully for many years for electric utilities delivering voltages greater than required for most electric launchers. A short-circuit fault near the terminals of such a generator could cause a substantial overcurrent to flow and would damage the machine. To prevent this, utility generators are deliberately designed so that the resistance seen by such a fault (the “subtransient reactance”) is high enough to limit the current to a value that does not cause serious damage. However, such machines can be modified to deliver a limited number of higher than normal current pulses. For example, in 1963, a 3-GVA machine was used for testing circuit breakers with asymmetrical currents of 300 kA at 16.5 kV (35).

A pulsed alternator deliberately designed with a low subtransient reactance to deliver safely a high current is an alternative to the HPG for an electric launcher. Instantaneous power of several GW is typically needed for electric launchers, usually for subsecond pulse durations. Such a machine has to operate under conditions so extreme that they would be acceptable only once or twice in the lifetime of a utility machine.

Pulsed iron-cored machines are too heavy for many electric launcher applications. Therefore, during the last few years, the United States has invested in the development of pulsed alternators that have disk (36) and drum topologies (37,38). These machines can be driven up to full speed in seconds to minutes and then discharged into the electric launcher. Multipolar geometries are used to ensure compactness so that the ac output current has to be rectified for launchers that require dc. The rectifier and switching assembly can be a significant part of the total system mass, volume, and cost. High rotational speeds enable small machine sizes to be achieved but necessitate the use of lightweight, high-strength rotor materials in which the conductors necessary to carry current are embedded. Carbon fiber composites are the modern structural choice. At high rotational speeds, aerodynamic losses become important, and it may be necessary to partially evacuate the machine interior to minimize heating.

One feature shared by all rotating machines is their capability to store more energy inertially in the rotor than required for each shot. The voltage, power, and energy storage requirements need to be evaluated for each particular launcher type and mission to give the optimum design. In some cases, the machine size and geometry may be determined by the electrical design (required output voltage or power), but in other cases the stored energy is the dominant feature. Decoupling the design for electrical power production from the energy storage design may be possible by using flywheels that are separate from the power generator.

Linear Magnetic Flux Compressors. Linear flux compressors have been suggested as an alternative to rotating machines and capacitors for driving electric launchers (39). By using large quantities of high explosives to drive conductors together to compress a magnetic field, current pulses greater than 250 MA have been achieved in microsecond pulses at low efficiencies (<1 to 10%).

Neither of these conditions is ideal for an electric launcher that might require pulses of a few milliseconds, especially because the use of explosives implies single-shot operation. Replacing explosives with a fuel/air combustion system has been suggested for multi-shot systems (40,41) but has not been tested. Flux compressor efficiencies are likely to fall to low levels as pulse lengths are extended, as a result of diffusion of the magnetic field into the conductors and correspondingly increased resistive losses.

Magnetohydrodynamic Generators. In a magnetohydrodynamic (MHD) generator, the flow of an electrically conducting gas (a “plasma”) through a magnetic field generates a voltage which, when extracted through electrodes in a generator channel, can transfer current to an external load without the intervention of any rotating element, such as a turbine or generator. A practical device consists of a plasma source and a channel, through which the plasma flows following acceleration in a supersonic nozzle, together with a magnet and electrodes embedded in the channel. The plasma is produced by reacting fuel and oxidant, together with “seeding” by a material of low ionization potential, such as a potassium or cesium compound. Early electric utility concepts also considered high-temperature, gas-cooled nuclear reactors cooled with cesium-seeded helium as the heat source. It is possible that a future space system could employ a similar Brayton or Rankine-cycle arrangement.

Considerable research was undertaken in the 1960s and 1970s on MHD devices for long-duration, low-current, utility power generation, but relatively little research has been done for pulse lengths of a few milliseconds. Repetitive, low-current (kA) pulses at a frequency of ~ 2 Hz and rise/fall times of tens of milliseconds were generated for pulses of a few seconds during operation of a liquid oxidizer/solid fuel rocket-driven generator (42). Currents as low as this would need to be transformed up for most electric launcher applications. Explosively driven, highly ionized argon systems have produced extremely high power ($> \text{GW}$) and high current ($> \text{MA}$) for tens of microseconds (43).

The claimed advantage of MHD generators is the relative simplicity of their “direct conversion” of thermal to electrical energy. However, the relatively low conductivity of most partially ionized plasmas limits the current that can be obtained from MHD generators to 10's kA, rather than MA desired for large electric launchers. One experiment to power a low-current rail gun with a MHD generator via energy storage in an inductive current transformer has been reported (44).

Inductors. Pulsed inductors have been used for intermediate energy storage and as a pulse compression component in HPG-powered rail-gun systems (11,45,46). Any low-voltage source including batteries, can be used to provide current for the inductor. The inductor can be regarded as a pressure vessel for mechanical design purposes; the stored magnetic energy is given by

$$E = B^2 / 2 \mu_0, \quad (11)$$

where B is the magnetic field strength (teslas) and $\mu_0 = 4\pi \times 10^{-7} \text{ H/m}$. A toroidal inductor is ideal for minimizing EM interference with external circuits caused by fringing magnetic fields. Shielded solenoids are a less expensive alternative that can also have low fringing fields.

For large EM launchers, the need to store and deliver tens of MJ at a few MA necessitates inductances L of a few μH . Using high currents, it is difficult to achieve system resistances R less than tens of $\mu\Omega$ with room temperature conductors, so that characteristic current decay times ($\tau = L/R$) are typically < 0.1 s. This requires discharging the inductor very quickly after charging, otherwise substantial current decay and losses occur. It is helpful to cool the inductor to cryogenic or even superconducting temperatures, so that R is reduced and τ increased. Cooling copper to liquid nitrogen temperatures (77 K) is beneficial in reducing resistance. Liquid hydrogen (20 K) with high purity (99.999% purity) aluminum is even better but requires careful handling. At liquid helium temperatures (4 K), many materials become truly superconducting, and large inductors have been made with NbTi and NbSn₃ conductors for nuclear fusion experiments. Liquid helium is unlikely to be acceptable on cost or logistic grounds for military applications but could be attractive for static ground-based systems that operate for long times with low losses.

The last decade has seen spectacular progress in high temperature superconductors based on copper oxide compounds: superconductivity has been demonstrated at temperatures well into the liquid N₂ range (over 130 K). So far, critical current densities are low, and present conductors are not yet candidates for high-field inductors. Nevertheless, there are promising developments in the development of high current density ($> 1 \text{ MA/cm}^2$), YBaCuO films on flexible substrates that could be used for inductors, and further improvements seem likely (47).

Batteries and Flywheels. The two options usually considered for energy storage are chemical batteries and flywheels. Batteries store large amounts of energy very compactly (a standard lead-acid automobile battery may contain 50 MJ) but cannot deliver high-power pulses. The recent international interest in developing electric vehicles for the consumer market is likely to lead to battery improvements. Unfortunately, electric vehicle requirements do not match the energy or pulse length needs of electric launchers. Modified batteries for high-power delivery generally use bipolar geometry, but even these are not well matched to most electric launcher requirements unless coupled with a multi-stage inductor (48). Lithium-ion battery technology has recently progressed rapidly, but other candidates are also possible. In all cases, batteries (and a similar system, the ultracapacitor) are based on electrochemical reactions that have characteristic voltages of 2–4 volts, so that multicell units are necessary to get voltages of 100–500 V. The voltages needed for electric launchers require multiple series/parallel strings or, more likely, electronic inverters to transform the battery dc output to high voltage dc.

Flywheel energy storage is another alternative. Very high energy storage densities can be achieved, especially at high peripheral rotor speeds. Oak Ridge National Laboratory has achieved peripheral speeds > 1400 m/s using carbon fiber composites (49). Less highly stressed versions of such flywheels are now in service commercially on a limited scale for passenger buses in Europe. There are significant differences between long-term energy storage and discharge for commercial applications versus the short pulses needed for electric launchers. The flywheel requires support subsystems for bearings, lubrication, and vacuum housing and needs to be coupled to a generator with a transformer/rectifier to produce high-voltage dc for launchers.

Auxiliaries. All of the power delivery and energy storage devices described here require auxiliary components. For example, switches are required to connect the pulsed power source to the electric launcher. These need to transfer high currents and to hold off high voltages. Options include solid-state devices, spark-gap switches, vacuum switches, and plasma-filled devices. Other auxiliaries depend on the specific power technology. For example, high-speed rotating generators require brushes and slip rings, vacuum enclosures, and cabling to connect the components together and to the gun breech. Operation of the entire launcher system requires a control system that includes the diagnostics necessary to monitor the total system health. This control system must integrate inputs from the prime power, pulsed energy and power system, launcher control system, and an overall mission control system.

Space Applications

First Stage/Orbiter Launch. A natural progression from the early work of Birkeland and Northrop was for low-velocity applications of launcher technology to merge with the development of linear induction motors. Linear motors are usually considered for continuous applications such as “people movers” in transit systems or for intercity trains where magnetic levitation can be employed in “maglev” systems. These motors operate continuously. A class of applications that has arisen recently is linear motors in amusement park rides in which short-term “launches” of a few seconds provide passengers the “thrill” of a high (or low)-g ride. An extension of this is to the launch of aircraft or other military equipment.

During World War II, a full-scale EM catapult called the “Electropult” was developed and built by the Westinghouse Corporation for the assisted launch of U.S. bombers from short runways on small mid-ocean atolls in the Pacific (50). Although successful, the project was finished too late to enter the war and was not pursued further. Similar efforts have been restarted in the last few years by the U.S. Navy to develop an EM Aircraft Launcher System (EMALS) to replace the steam catapult on an aircraft carrier (51). Depending on the results during the next few years, it is possible that future aircraft catapults could be of this type. The masses launched in these systems are 5–50 tons, and the required velocity is usually 25–100 m/s, so that launch energies of 50–100 MJ are typical. The challenge in this case is to accomplish the launch in the limited distance onboard the aircraft carrier, which is about 100 meters. The aircraft is therefore subjected to an acceleration of about 3 gs. The most important system concern is reliability; failure rates need to be less than 1 in 10 million to be acceptable for naval aviation.

The NASA Space Flight Center in Huntsville is currently testing a small-scale system of this type that could form the basis for a future replacement of the first stage for an Earth-to-space orbiter launch system (52). This could eliminate the first stage of a multistage missile system by boosting the orbiter to the speed at which the later stages can be ignited for further propulsion. A system of this type, known as *Maglifter*, was proposed in 1994 (53). A similar proposal, made recently by the Russians, suggested the use of MHD generators to power the

launcher (54). Based on the technical success of experimental maglev trains in Japan and Germany, the successful development of such a system seems quite feasible.

Direct Launch to Space. It is remarkable that one application that was apparently considered by the early developers of EM technology was that of launching manned or unmanned objects directly into space.⁷ The interest was based on the idea that, in principle, EM forces can act at very high speeds, so that there should not be any *fundamental* physical barriers to their operation. Though this is largely true, some physical effects have been encountered at speeds in excess of 3 km/s that are not yet fully understood, and the velocities required for a direct Earth-to-space launch have not yet been reached except with gram-sized projectiles.⁸

Two concepts have been studied—the coil-based inductive system and the rail-gun system. The earliest versions of coil launchers, referred to as “mass drivers,” were studied at Princeton University by O’Neill and his colleagues (55,56) and at MIT in the late 1970s and early 1980s. Only relatively low velocities of (<100 m/s) were achieved, but because O’Neill’s suggestion was to use the mass drivers for escape from the Moon or to move objects from one orbit to another, very high velocities were not necessary. What was important was the use of renewable energy—so that solar panels were proposed. Scientists at the Sandia National Laboratories, where velocities up to 1 km/s have been achieved with a coil system (25), have suggested launch-to-space using a coil gun (57), but the required muzzle velocity of 6 km/s has not been achieved. Rail guns have been much more successful in demonstrating high velocities (up to 4.3 km/s with 0.6 kg and ~ 7 km/s with gram-sized projectiles) and high energies (up to 9 MJ), and this has led to a number of detailed studies of possible space launch (58–65).

Very high energies and power levels are required to accelerate large masses into space. Launching a 1000-kg aeroshell of the type illustrated in Fig. 7 requires an armature and sabot system that may make the total mass 1250 kg (see Fig. 8). At a launch velocity of 7.5 km/s, this represents a muzzle energy of 35 GJ. One suggested approach would use a staged transiently augmented rail-gun (STAR) launcher of the type illustrated in Fig. 6 (65).

This concept offers very high efficiency because the energy needed for the launch is transferred from section to section along the launcher during acceleration. With a launcher efficiency of 80%, the energy input for the launch of this package would be 44 GJ. The cost of the electricity for such a launch would be negligible—less than \$1000. A substantial capital investment would be required to build, install, and operate such a facility, including a dedicated power supply or link to the utility network. For a moderate peak acceleration (by gun standards) of 30,000 g’s, the launcher would be some 230 meters long while for 2000 g’s,

⁷There is a report that the Russian, Rynin, stated in 1929 that “the Austrian, Heft, in 1895 proposed a solenoid gun for launching interplanetary spaceships at speeds exceeding the 11.8 km/s needed to reach interplanetary space.”

⁸Launch from the Moon or from an asteroid would require a much lower velocity and may be within the reach of present laboratory velocities. However, there is no requirement now for a system with such capabilities.

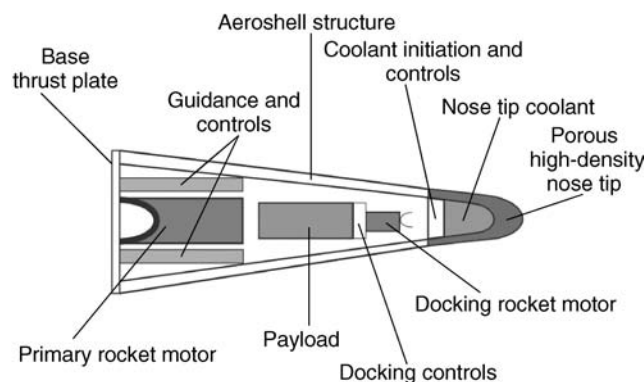


Figure 7. A flight body concept. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

it would be about 1 mile long.⁹ One recent estimate of the cost of such a system is about \$1.4B (65) (see Table 6). A proposal for an induction launcher system also estimated a capital cost of \$1–\$2B. These costs are competitive with conventional launch methods and are similar to the cost of a single Shuttle-Orbiter vehicle.

System needs have not yet been fully evaluated. One of the most important issues (for an Earth-based system) is the “inverse reentry” problem of getting a gun-launched projectile up through the atmosphere without substantial ablation. Techniques for dealing with this, such as sacrificial ablative coatings or gas/vapor nose-tip cooling, appear feasible at first analysis. A mitigating factor is that the projectile should transit the atmosphere in only a few seconds. Launching from an equatorial site at high altitudes, as shown conceptually in Fig. 9 (66), would be doubly beneficial because the atmospheric density would be lower and there is about a 400 m/s velocity contribution from Earth’s rotation.

Part of the total mass to be launched will be a rocket motor that is needed to circularize the orbit after reaching apogee (see Fig. 8). In addition, some divert capability may also be needed to ensure that the payload can reach a collection point in space, such as the International Space Station. As a result, only 20 to 30% of the aeroshell mass will be useful payload. Hence, if tens or hundreds of tons of supplies were to be launched in this way, many launches would be required. For example, to supply 100 tons of materiel into orbit in payload packages of 250 kg would require 400 launches per year, which is equivalent to two per day for 200 working days. Launching does not seem an impossible task (a power station of only about 10 MWe would be required), but management of the packages in orbit will require an infrastructure that does not presently exist.

EM Spacecraft Propulsion. To reduce the mass of fuel that has to be carried by a rocket, it is beneficial to increase the exhaust velocity of the propellant. Chemical rockets generally operate with a specific impulse (I_{SP}) in the range from about 250 to 400 seconds. To increase this, other techniques need to be developed. These include ion thrusters which have an $I_{SP} \sim 2000$ to 3000 s,

⁹Of course, only rugged materials such as fuel, water, food, or engineering supplies could be launched at such accelerations.

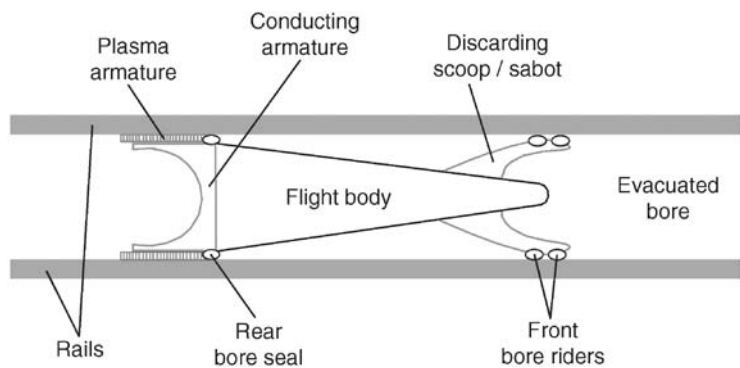


Figure 8. The launch package concept. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

plasma thrusters, and even EM launchers. All of these are applicable only for use in space, not for launch into space from the surface of a planet.

Two studies of EM launchers for spacecraft propulsion were published in the early 1980s. Snow, et al. evaluated an induction launcher (known as the “mass driver”) as a reaction engine for LEO to GEO orbit transfers (67). The goal was an I_{SP} of 500–1000 s, which requires an exhaust velocity of 5 to 10 km/s. As discussed earlier, induction launcher performance is far from this at present. Bauer et al. (68) proposed to use a rail gun that launched pellets of an unspecified material at velocities of 5 to 20 km/s, corresponding to an I_{SP} of 500–2000 s. Apart from the failure of induction or rail launchers to reach the required velocities, a major concern with these concepts is the hazard created for other spacecraft, such as commercial satellites in near-Earth orbits, from ejecting high-velocity projectiles. In principle, these techniques could be considered for other missions, but there are competing technologies that are likely to be more attractive.

U.S. and International Programs

The best resources to consult for a review of EM launcher developments are the *Proceedings of the EM Launcher Symposia* that have been held every 2 or 3 years since 1982. These conferences have been published in archival form in special editions of the *IEEE Transactions on Magnetics* (69–78). Some papers can also

Table 6. Capital Cost Estimate for an Earth to Space EM Launcher

Component	Cost per unit	No. units	Total component cost
Ac generators	\$3 M	100	\$300 M
Switching	\$3 M	100	\$300 M
Barrel	\$0.3 M/m	1600 m	\$480 M
Busbars and structure			\$100 M
Civil engineering			\$250 M
Total			\$1410 M

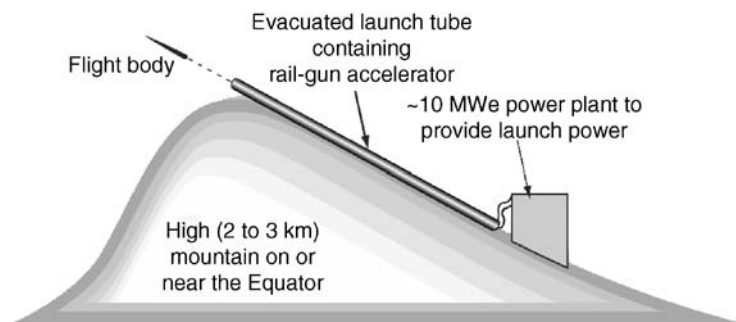


Figure 9. The preferred launch site layout.

be found in the nonarchival, but widely circulated, *Proceedings of the Pulsed Power Conference*, which is also now sponsored by the IEEE (79). Proceedings of specialized journal issues or conference in different countries are also available occasionally (80). Because of the defense-related nature of much of this work, as well as the desire of industrial companies to protect their proprietary concepts, it seems likely that much R&D has occurred that has not been openly published. This overview is based on information from open sources.

United States. Although one or two isolated studies were undertaken in the late 1950s and 1960s, it was not until the mid-1970s that significant efforts began to develop in the United States. Following the approach pioneered at the Australian National University, several of these programs used homopolar generators as the energy and power source. These include facilities at Picatinny Arsenal (30 MJ), the University of Texas (60 MJ), and Eglin Air Force base (13 MJ). Large facilities were built with capacitors at Picatinny Arsenal (52 MJ) and Westinghouse (32 MJ). One facility installed by the USAF at Eglin used thousands of automotive batteries to charge an inductor that powered a rail gun. A larger number of small facilities were also used for R&D, based mostly on capacitors in independent modular arrangements. The largest rail guns used were 90-mm bore diameter and up to 8 meters long. Muzzle energies achieved with these rail guns were close to 9 MJ. In the last few years, financial strictures have forced rationalization of these facilities, and most have been dismantled or “mothballed.” Coil-gun research has been more sporadic; the last remaining organization that has interest in this area is Sandia National Laboratories, although this effort is now not active.

Europe. Efforts on EM launchers and related technology developments are being, or have been, undertaken in recent years in Britain, Germany, France, Italy, Sweden, Israel, the Netherlands, and Switzerland. Experimental facilities have been established in several locations; the largest are at Kirkcudbright in Scotland (33 MJ), Unterluss in Germany (30 MJ), and St. Louis in France (10 MJ). Smaller facilities are available in the other countries. Several of these facilities use power supplies based on capacitors that were supplied by U.S. manufacturers¹⁰. TNO in the Netherlands also used a refurbished U.S. homopolar generator, and Magnet Motor GmbH in Germany is developing flywheel

¹⁰In some cases, through cooperative arrangements with the U.S. Department of Defense.

systems with permanent magnet generators that can serve as MW power sources to charge superconducting, inductive, pulsed GW systems. Research in Britain and France has been on rail guns and electro thermal chemical (ETC) guns. Germany, Israel, and Sweden have concentrated solely on ETC gun technology. **The Former Soviet Union.** During the latter years of the Soviet Union, considerable R&D was undertaken on EM launchers for military applications, although relatively little was published in the West. Russia and the Ukraine were the primary developers, and several significant systems were built. Perhaps because capacitor technology is less well advanced in these countries, greater use seems to have been made of rotating machines. One notable Ukrainian system that was installed in a transportable trailer used four homopolar generators, powered by helicopter gas turbines, that had 60 MJ stored energy. Other rotating machine system designs were for even larger energies. Based on published reports, there appears to have been more emphasis on achieving very high velocities in Russian research than in the United States, with several reports of 6 or 7 km/s with small pellets of a few grams (81) and possibly up to 10 km/s with milligram masses.

East. One of the early efforts in developing rail guns was undertaken at the Ministry of Defense in Australia in the mid-1960s. The availability of a very large (550 MJ) homopolar generator at the Australian National University (ANU) and interest in the effects of micrometeorite damage on spacecraft led to a larger effort at ANU in the early 1970s. Experiments were undertaken at velocities close to 6 km/s with gram-sized projectiles. This work ended after the two main researchers moved to the United States in the late 1970s.

During the late 1980s and early 1990s, several programs on EM launchers were funded at a modest level by industrial companies in Japan. In most cases, capacitor banks were used that were purchased from U.S. companies. Several innovative solutions were developed with modest investments. In the related area of high-speed transport, Japan, like Germany, has strong research programs and a national commitment to improved high-speed transit technology. This could form a good foundation for induction launcher research for higher velocities.

Serious Chinese studies started in the late 1980s, but publication occurred only in Western conferences and journals in the mid-1990s. A wide range of technical interests and level of technical sophistication are evident in recent publications, that show interest in both EM and ETC technology. The work builds on what has been undertaken in the West but with its own variations. Summaries of Western work have been prepared by Chinese authors and are available as a foundation for further research (82).

South Korea began to take an interest in this technology in the mid-1990s, and there are now several technical organizations involved with their Ministry of Defense in this research, especially on ETC guns. A Korean Conference on this technology was held in 1997.

Future Possibilities

Despite the many potential advantages of EM propulsion enumerated by Fauchon-Villeplee more than 80 years ago, military applications of rail guns are not

yet guaranteed. Further developments during the next few years will be important in showing what can be accomplished and where this technology should be best applied.

Much has been done in projectile, barrel, and power supply development for rail guns; velocities of 2000 to 3000 m/s are routinely reached with kilogram-sized projectiles and up to 7000 m/s with gram-sized masses. It seems likely that a research program on the technical challenges of rail guns for very high velocity applications, such as launch-to-space, would be most interesting and rewarding.

In contrast, much more has yet to be demonstrated before inductive launchers are likely to prove useful for high-velocity applications; the maximum velocity achieved so far is about 1000 m/s. The first use of induction technology seems likely to be for low-velocity applications; such as an EM catapult for the Navy. It is feasible that this could be followed by the replacement of the first stage of a launch system for an orbiter vehicle using the Maglifter (58) or a similar concept.

BIBLIOGRAPHY

1. Page, C.G. New electromagnetic engine. *J. Sci. Arts (US)* 49: 131–135 (1845).
2. Egeland, A. Birkeland's electromagnetic gun: A historical review. *IEEE Trans. Plasma Sci.* 17 (2): 73–82 (1989).
3. Fauchon-Villeplee, A.L.O. *Canons electriques*. Berger-Levrault, Nancy, Paris, Strasbourg, 1920.
4. Korol'kov, A.L. Long range electrical gun: Equipment and supplies of the red army. *Tekhnika i Snabzheniye Krasnoy Armii* 37 (68): 1–6, 8 (20 August 1923).
5. MacLaren, F.B. U.S. Pat. 1,384,769, 19 July 1921.
6. Pseudoman, A. *Zero to Eighty*, Scientific Publishing, Princeton University Press, Princeton, NJ, 1937.
7. Liebhafsky, H.A. Gesellschaft ur Geratebau. Combined Intelligence Objectives Subcommittee Rept., 14 June 1945.
8. Pollack, M., and L.W. Matsch. Electric Gun and Power Source. Armour Research Foundation Report No.3, Project 15-391E. 1 May 1947.
9. Mannal, C., and Y.A. Yoler. Acceleration of masses to hypervelocities by electromagnetic means. *Proc. 2nd Hypervelocity Impact Effects Symp.*, December 1957, Vol. 1.
10. Brast, D.E. and D.R. Sawle. Feasibility Study for Development of a Hypervelocity Gun. MB Associates Report MB-R-65/40. May 1965. (NASA Contract NAS 8-11204), pp. 75–79.
11. Barber, J.P. The Acceleration of Macroparticles and a Hypervelocity Electromagnetic Launcher. Ph.D. Thesis EP-T12, The Australian National University, Canberra, March 1972.
12. McNab, I.R., F. LeVine, and M. Aponte. Experiments with the Green Farm electric gun facility. *IEEE Trans. Magn.* 31 (1): 338–343 (1995).
13. Kerrisk, J.F. Current Distribution and Inductance Calculations for Rail Gun Conductors. LANL Report LA-9092-MS, UC-34, 1981.
14. Homan, C.G., C.E. Cummings, and C.M. Fowler. Superconducting Augmented Rail Gun (SARG). *IEEE Trans. Magn.* 22 (6): 1527–1531 (1986).
15. Katsuki, S., et al. Behaviors of plasma armature in the augmented rail gun using a permanent magnet. *IEEE Trans. Magn.* 31 (1): 183–188 (1995).
16. Fikse, D.A., J.L. Wu, and Y.C. Thio. The ELF-I Augmented Electromagnetic Launcher. *IEEE Trans. Magn.* 20 (2): 287–290 (1984).

17. Werst, M.D., T.J. Hotz, J.R. Kitzmiller, C.E. Penney, and R.M. Telander. Testing of the cannon caliber rapid fire railgun. *IEEE Trans. Magn.* 33 (1): 613–618 (1997).
18. Marshall, R.A. The TAERF scientific railgun theoretical performance. *IEEE Trans. Magn.* 18 (1): 11–15 (1982).
19. Parker, J.V. Electromagnetic projectile acceleration utilizing distributed energy sources. *J. Appl. Phys.* 53 (10): 6710–6723 (1982).
20. Moldenhauer, J.G., and G.E. Hauze. Experimental demonstration of an N-Turn EML. *IEEE Trans. Magn.* 20 (2): 283–286 (1984).
21. Poltanov, A.E., A.K. Kondratenko, A.P. Glinov, and V.N. Ryndin. Multi-turn railguns: Concept analysis and experimental results. *IEEE Trans. Magn.* 37 (1): 457–461 (2001).
22. Poltanov, A.E., A.P. Glinov, A.K. Kondratenko, and V.N. Ryndin. Use of multi-turn railguns as high speed limiters of short circuit current for large power plants. *IEEE Trans. Magn.* 37 (1): 229–231 (2002).
23. McNab, I.R. The STAR railgun concept. *IEEE Trans. Magn.* 35 (1): 432–436 (1999).
24. Asano, T., Y. Sakai, M. Oshikiri, K. Inoue, and H. Maeda. Development of long-pulsed high-field magnets. *Jpn. J. Appl. Phys.* 32: 1027–1029 (1993).
25. Turman, B. Long range naval coilgun technology. Briefing. 21 February 2001.
26. Petrov, S.R. Russian interests in coilguns. Personal communication to Z. Zibar, January and April 1994.
27. See, for example, Chapter V of Non-US Electrodynamic Launcher Research and Development, Ed. J.V. Parker, et al. SAIC FASAC Report, November 1994.
28. Bondaletov, V.N., and Ye.N. Ivanov. Ultrahigh axial acceleration of conducting rings. *Sov. Tech. Phys. J.* 22 (2): 232–234 (1977).
29. Augsburgers, B., B. Smith, I.R. McNab, Y.G. Chen, D. Edwards, S. Gilbert, G. Savell, M. Robinson, and P. Chapman. DRA 500 kJ multi-module capacitor bank”. *IEEE Trans. Magn.* 31 (1): 10–15 (1995).
30. Augsburgers, B., B. Smith, I.R. McNab, Y.G. Chen, D. Hewkin, K. Vance, and C. Disley. Royal Ordnance 2.4 MJ multi-module capacitor bank. *IEEE Trans. Magn.* 31 (1): 16–21 (1995).
31. Grater, G.F., F.W. McDougall, M. Hudis, and X.H. Yang. Characterization of high energy density capacitors under projected US Navy ETC gun operating conditions. *10th IEEE Int. Pulsed Power Conf.* Albuquerque, NM, July 10–13, 1995, Paper 15.2.
32. Winsor, P., M. Hudis, and K. Slenes. Pulse power capability of high energy density capacitor-based on a new dielectric material. *12th IEEE Pulsed Power Conf.*, Monterey, CA, June 27–30, 1999.
33. Sarjeant, W.J., B.D. Goodell, S.L. Langlie, and K.C. Pan. Technical tradeoffs for downsizing ETC power systems. *IEEE Trans. Magn.* 29 (1): 1054–1059 (1993).
34. McNab, I.R. Homopolar generators for electric guns. *IEEE Trans. Magn.* 33 (1): 461–467 (1997).
35. Kilgore, L.A., E.J. Hill, and C. Flick. A new three-million kVA short-circuit generator. *IEEE Trans. Pas.* 82: 442–446 (1963).
36. Curtiss, D.H., P.P. Mongeau, and R.L. Puterbaugh. Advanced composite flywheel structural design for a pulsed disk alternator. *IEEE Trans. Magn.* 31 (1): 26–31 (1995).
37. Kitzmiller, J.R., R.W. Faidley, R.L. Fuller, R.N. Hedifen, and R.F. Thelan. Manufacturing and testing of an air-core compulsator driven 0.60 caliber railgun system. *IEEE Trans. Magn.* 29 (1): 441–446 (1993).
38. Walls, W.A., S.B. Pratap, W.G. Brinkman, K.G. Cook, J.D. Herbst, S.M. Manifold, B.M. Rech, R.F. Thelan, and R.C. Thompson. A field based, self-excited compulsator power supply for a 9 MJ railgun demonstrator. *IEEE Trans. Magn.* 27 (1): 335–340 (1991).
39. Marshall, R.A. A reusable inverse railgun magnetic flux compression generator to suit the Earth-to-space rail launcher. *IEEE Trans. Magn.* 20 (2): 223–226 (1984).

40. Mongeau, P. Combustion driven pulsed linear generators for electric gun applications. *IEEE Trans. Magn.* 33 (1): 468–473 (1997).
41. Goldman, E.B., F. Davies, K.J. Bickford, E. Smith, P. Williams and M. Leone. Development of a flux compression power unit for millisecond ETC pulsed power applications. *IEEE Trans. Magn.* 35 (1): 340–345 (1999).
42. Demetriades, S.T., and C.D. Maxwell. Pulsed operation of a combustion MHD generator. *VIII IEEE Int. Pulsed Power Conf.* 1991, pp. 457–460.
43. Baum, D., and S.P. Gill. Pulsed electrical power studied. *Aviation Week and Space Technol.* April 27, 1981.
44. Babakov, Y.P., A.V. Plekhanov, and V.B. Zheleznyi. Range and railgun development results at LS&PA SOYUZ. *IEEE Trans. Magn.* 31 (1): 259–262 (1995).
45. Deis, D.W., and I.R. McNab. A laboratory demonstration electromagnetic launcher. *IEEE Trans. Magn.* 18 (1): 16–22 (1982).
46. Gully, J.H., D.J. Hildebrand, and W.F. Weldon. Balcones homopolar generator power supply. *IEEE Trans. Magn.* 25 (1): 210–218 (1989).
47. Lubkin, G.B. Power applications of high-temperature superconductors. *Phys. Today* 48–51 (March 1996).
48. Thomas, C.A. Inductor-based pulsed power. University of Texas paper, IAT, Austin, Texas, August 2001.
49. Gully, J.H. Power supply technology for electric guns. *IEEE Trans. Magn.* 27 (1): 329–334 (1991).
50. Jones, M.F. Launching aircraft electrically. *Aviation* 45 (10): 62–65 (1946).
51. Doyle, M.R., D.J. Samuel, T. Conway, and R.R. Klimowski. Electromagnetic aircraft launch system—EMALS. *IEEE Trans. Magn.* 31 (1): 528–533 (1995).
52. Wolcott, B. Induction for the birds. *Mech. Eng.* 66–70 (2000).
53. Mankins, J.C. The Maglifter: An advanced concept using electromagnetic propulsion in reducing the cost of space launch. *30th Joint Propulsion Conf.* June 27–29, 1994, AIAA Paper 94-2726.
54. Batenin, V.M., et al. Advanced reusable space transportation system for horizontal launch of Air-space plane powered by pulsed MHD generators. AIAA Paper 2001-0495.
55. Chilton, F. Mass driver theory and history. May 9–12 1977, AIAA Paper 77-533.
56. O'Neill, G. The High Frontier: Human Colonies in Space. William Morrow, New York, 1977.
57. Lipinski, R.J., et al. Space applications for contactless coilguns. IEEE Paper 0018-9464/93, pp. 691–695.
58. Hawke, R.S., A.L. Brooks, C.M. Fowler, and D.R. Peterson. Electromagnetic railgun launchers: Space propulsion applications. April 21–23, 1981, AIAA Paper 81-0751.
59. Rice, E.E., L.A. Miller, R.A. Marshall, and W.R. Kerslake. Feasibility of an Earth-to-space rail launcher system. *Int. Astronaut. Fed. Conf.* Paris, France, October 1, 1982, Paper IAF-82-46.
60. Palmer, M., and A.E. Dabiri. Electromagnetic space launch: A re-evaluation in light of current technology and launch needs and feasibility of a near-term demonstration. *IEEE Trans. Magn.* 25 (1): 393–399 (1989).
61. Fair, H.D., P. Coose, C.P. Meinel, and D.A. Tidman. Electromagnetic Earth-to-space launch. *IEEE Trans. Magn.* 25 (1): 9–15 (1989).
62. Schroeder, J.M., J.R. Gully, and M.D. Driga. Electromagnetic launchers for space applications. *IEEE Trans. Magn.* 25 (1): 504–507 (1989).
63. Palmer, M.R. Motivation for a near earth gun launch to space demonstration and a variable inductance power supply concept to minimize initial demonstration costs. *IEEE Trans. Magn.* 29 (1): 478–483 (1993).
64. Brown, J.L., et al. Earth-to-space railgun launcher. *IEEE Trans. Magn.* 29 (1): 373–378 (1993).

65. McNab, I.R. Launch to space with an electromagnetic railgun. *11th Electromagnetic Launch Symp.* Saint Louis, France, May 14–17, 2002. To be published in the *IEEE Trans. Magn.* 39(1): (2003).
66. Gasner, D.R. Private communication. May 2001.
67. Snow, W.R., and R.S. Dunbar. Mass driver reaction engine characteristics and performance in earth orbital transfer mission. *IEEE Trans. Magn.* 18 (1): 176–189 (1982).
68. Bauer, D.P., J.P. Barber, and H.F. Swift. Application of electromagnetic accelerators to space propulsion. *IEEE Trans. Magn.* 18 (1): 170–175 (1982).
69. 1980 conference on electromagnetic guns and launchers. *IEEE Trans. Magn.* 18 (1): 3–216 (1982).
70. 2nd symposium on electromagnetic launch technology. *IEEE Trans. Magn.* 20 (2): 197–411 (1984).
71. 3rd symposium on electromagnetic launch technology. *IEEE Trans. Magn.* 22 (6): 1377–1832 (1986).
72. 4th symposium on electromagnetic launch technology. *IEEE Trans. Magn.* 25 (1): 6–670 (1989).
73. 5th symposium on electromagnetic launcher technology. *IEEE Trans. Magn.* 27 (1): 4–660 (1991).
74. 6th symposium on electromagnetic launcher technology. *IEEE Trans. Magn.* 29 (1): 337–1224 (1993).
75. 7th symposium on electromagnetic launch technology. *IEEE Trans. Magn.* 31 (1): 6–775 (1995).
76. 8th symposium on electromagnetic launch technology. *IEEE Trans. Magn.* 33 (1): 6–657 (1997).
77. 9th symposium on electromagnetic launch technology. *IEEE Trans. Magn.* 35 (1): 5–489 (1999).
78. 10th electromagnetic launch technology symposium. *IEEE Trans. Magn.* 37 (1): 5–511 (2001).
79. Stallings, C., and H. Kirbie (eds.). *12th IEEE Int. Pulsed Power Conf.* Monterey, CA, IEEE Catalog Number 99CH36358 (1999).
80. Special Issue on Electromagnetic Launchers. *IEEE Trans. Plasma Sci.* 17 (3): 349–564 (1989).
81. Dobyshevskii, E., et al. Experiments on simple railguns with the compacted plasma armature. *IEEE Trans. Magn.* 31 (1): 295–298 (1995).
82. Ying, W., and X. Feng (eds.). *Principles of the Electric Gun.* Chinese Defense Industry Press, LN285-95, Beijing, China, 1995.

IAN R. MCNAB

The Institute for Advanced Technology
The University of Texas at Austin
Austin, Texas

EVOLUTION OF U.S. EXPENDABLE LAUNCH VEHICLES

Beginnings

Robert H. Goddard was a Professor of Physics at Clark University in Worcester, Massachusetts, a post he held for 26 years from 1919 until his death in 1945.

During these years, Goddard performed a classic series of experiments in which he, almost single-handedly, worked out the basic elements of liquid-fueled rocket technology. Goddard was interested in using rockets for space flight from the very beginning. He felt that for this purpose, liquid fuels, specifically, kerosene or gasoline and liquid oxygen would be best. This fuel would be more efficient, that is, it has a higher specific impulse, and it has the advantage that by closing and opening valves, the rocket engine could be stopped and restarted. This latter point was especially important because it gave liquid-fueled rockets a truly decisive advantage over those that operated with solid fuel. In solid-fueled rockets, once the motor is started, the rocket is committed, and the engine cannot be turned off.

In 1919, Goddard published a paper titled "A Method for Reaching Extreme Altitudes" in the *Journal of the Smithsonian Institution* in which he outlined his plans (1). These came to fruition on 16 March 1926, when a liquid-fueled rocket built by Goddard performed the first successful liftoff and flight and reached an altitude of about 80 feet (Fig. 1). This experiment was carried out in an open field on a farm near Auburn, Massachusetts. Goddard realized that this was not the best place for these experiments. He took advantage of an offer from Daniel Guggenheim, the heir to a huge mining fortune and an enthusiastic supporter of American aviation, to support the continuation of his experiments and to set him up on a large ranch near Roswell, New Mexico. It was here that Goddard built and flew the first gyroscopically stabilized liquid-fueled rocket on 31 May 1935; it reached an altitude of more than 7000 feet.

Goddard's remarkable success was unfortunately not pursued. Because of the depression, Guggenheim could no longer support Goddard's work, and eventually, the work at the Roswell ranch was abandoned. The U.S. military showed no interest in Goddard's work. When the Second World War started, the military did initiate a vigorous effort to develop solid-fueled rockets for use as airborne and ground-based weapons, in which Goddard participated. Goddard's untimely death in 1945 precluded his participation in the rapid development of large liquid-fueled rockets in the United States following World War II.

While all of this was happening in the United States, a group of Germans was also interested in developing liquid-fueled rockets for the same reasons that motivated Goddard. The intellectual leader of this group was Professor Hermann Oberth, a Romanian-German, who wrote the first book on planetary exploration that was published in 1923, "Die Rakete zu den Planetenraumen" (2). He advocated using rocket propulsion. In this book, he worked out the necessary mathematics to accomplish this objective and also elaborated on the infrastructure that would have to be built. (Both Goddard and Oberth were unaware of the work of Konstantin E. Tsiolkovski, a Russian mathematics professor who had worked out the rocket equation and had speculated on the use of rockets for space travel 20 years before Goddard and Oberth wrote about their work.) Oberth's book attracted the attention of a small group of German scientists and engineers who had a great interest in space travel and rockets. Along with Oberth, these included Rudolf Nebel, Klaus Riedel, Willy Ley, and Max Valier. On 5 July 1927, these people and a few others founded the "Verein für Raumschiffahrt," the "Society for Space Travel." By 1929, the Society had almost 900 members, one was a 17-year-old high school senior named Wernher von Braun. Von Braun was a



Figure 1. Robert Goddard with the liquid-fueled rocket that first flew on March 16 1926, (photo courtesy NASA Goddard Space Flight Center). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

member of a prominent and affluent Prussian family. His father served as Minister of Agriculture and also as Minister of Education under the democratic German Weimar Republic.

The Society acquired a small tract of land a few miles north of Berlin that was a German army ammunition storage depot and could be used for rocket flight experiments. The area was named the "Rocket Flight Area" (Raketenflugplatz, in German), and the Society began to conduct flight experiments. The technical leaders of the Society were convinced—as was Goddard—that only liquid-fueled rockets were practical for space travel. They knew about Goddard's work, and they began to build and test liquid-fueled rockets. The first test models were called "Mirak" which stood for the German "Minimum Rakete." They were small test rockets designed only for static tests. They also built flight models called "Repulsors" which looked very much like Goddard's first rocket. One of these flew for the first time on 10 May 1931 and reached an altitude of 18 meters. The rockets following this event were somewhat more sophisticated, but none reached the technical complexity of Goddard's gyrostabilized rocket.

Both Oberth and Nebel had talked and written about military applications of rockets. There was interest in rockets in Germany because the Versailles Treaty that ended World War I contained provisions that seriously limited German artillery development. In 1929, the Army Ordnance Department established a small group to study the possible military applications of rockets. A year later, a young German artillery captain, Walter Dornberger, was named to head this group. Dornberger was aware of the experiments conducted by Oberth, Nebel, and the others at the Raketenflugplatz. He decided to help the group by providing modest financial support. He also quickly came to the conclusion that the Raketenflugplatz site was not adequate, and he offered the facilities of the German artillery range at Kummersdorf for the rocket experiments. These moves accelerated technical progress, but they also weakened the Society because several of the people who worked on the Society's rockets joined Dornberger's group. Wernher von Braun, one of them, started to work at Kummersdorf on 1 October, 1932.

When Adolf Hitler and the Nazi party came to power in January 1933, things changed drastically. The civilian organizations interested in rocketry disappeared, and the work on military rockets was substantially accelerated. During the mid-1930s, the Germans developed a well-organized, systematic research program on liquid-fueled rockets. The first of these rockets, called the A-2, was successfully launched in December 1934 and reached an altitude of 2000 meters, about 6000 feet. It was not as sophisticated as Goddard's gyrostabilized rocket that was launched a few months later, but the A-2 did work. Wernher von Braun had acquired a Ph.D. and was the technical director of the Kummersdorf enterprise reporting directly to Dornberger. In any event, the Germans had now roughly caught up to where Goddard was when he was forced to quit.

By the mid-1930s, the Germans had made enough progress that they decided to develop a test site which would make it possible for them to test large rocket vehicles. They chose an isolated region on the north German Baltic coast where the River Peene reached the sea. The place was called Peenemünde and by 1939, test operations were started there. The A-2 was followed by the A-3 and the A-4, the latter was successfully flown on 3 October, 1942, almost a decade to the

day that von Braun joined Dornberger's unit. The A-4 was the prototype of the V-2, the first long-range weapon based on rocket technology (Fig. 2). It could throw a payload of 1 metric ton about 250 miles. The rocket engine of the A-4 developed a thrust of 60,000 pounds. The rocket burned ethanol as a fuel and used liquid oxygen as an oxidizer. Due to the success of the A-4 launch and the loss of the Battle of Britain in 1941, Hitler decided, on 7 July 1943, to put the development and production of the V-2 missiles at the highest level of priority.

Because of good air reconnaissance and brilliant photo interpretation, the British were aware that new and dangerous weapons were being developed and tested at Peenemunde. They mounted a massive air raid on the test site, which was partially successful. The Germans realized that they would have to disperse their test and production facilities because more allied air raids attacking the Peenemunde complex were inevitable. The Germans built a massive underground facility in central Germany called the "Mittlewerk" for producing V-2 missiles. They also built a test site in Poland which was out of range of long range bombers based in England. A little more than 6,250 V-2s were produced. About 3700 were fired at targets in Great Britain and on the continent, of which about 700 failed! About 2000 were in storage at the end of the war, 300 were expended in tests, and 250 were taken to the United States at the end of the war. Although



Figure 2. A V-2 missile taking off from the launch pad (photo courtesy NASA Marshall Space Flight Center). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the outcome of the war was not influenced by the use of large numbers of V-2 rockets, the creation of the V-2 missile and its successful operation was still a technical achievement of the first rank. It has been said that if the Manhattan Project to produce nuclear weapons was the technical “tour de force” of the United States during World War II, then the development of the V-2 played the same role in Germany.

The End of World War II

The V-2 rocket team that Wernher von Braun organized at Peenemunde was dispersed at the end of World War II. Wernher von Braun and most of the 100 or so senior technical people were housed in the small mountain village of Oberjoch on the Austrian-German border. They wanted to surrender to the U.S. Army, which had some units in the neighborhood, but they did not know how to make contact. At the same time, in 1945, the U.S. Army realized that it would be important to capture the Germans who had the knowledge to develop and build long-range missiles. An operation dubbed “Paperclip” was initiated to locate and detain the people who had this expertise. It was headed by U.S. Army Colonel Holger N. Toftoy. Eventually, the von Braun group surrendered to a U.S. Army unit and was taken into custody. Other veterans of the V-2 program wound up as prisoners of the Soviets. There were a few who actually volunteered to go to work for the Soviets.

Wernher von Braun, Walter Dornberger, and other leaders of the V-2 program arrived at Newcastle Air Base in Wilmington, Delaware on 19 September 1945. They were now under contract to the U.S. Army and were going to work on missile development. They were taken to Fort Bliss near El Paso, Texas, the U.S. Army’s Center for Anti Aircraft Artillery. Fort Bliss is located at the southern end of the White Sands Missile Test Range in New Mexico. It would be the job of von Braun and his colleagues to launch some of the V-2s brought over from Germany, to perform research, and to train Americans in the art and science of rocketry. In doing so, a large number of V-2 rockets was used; some were modified to measure the upper atmosphere and to photograph Earth, the Sun, and other astronomical objects. All of these activities were very important to the future of U.S. ballistic missile and space flight programs (3).

At the same time that the Germans were working for the U.S. Army, a group at Douglas Aircraft Company in Santa Monica, California, published a far-reaching document in 1946. The group was led by some outstanding scientists and engineers including Francis Clauser, David Griggs, and Louis Ridenour among others; they later left Douglas to form the RAND Corporation. The title of their work was “Preliminary Design of an Earth-Orbiting Spaceship” (4). The group anticipated that large rockets, based on extensions of the V-2 designs, would be developed both by the United States and the Soviet Union to carry nuclear weapons at ranges from 8,000 to 10,000 miles. Such rockets could be modified to place significant payloads into Earth orbit and to other places in the solar system. Thus, the means to realize the early ideas of Tsiolkovski, Goddard, and Oberth would soon be available, and therefore, it was time to make some detailed calculations about spaceflight. The remarkable thing about

“Preliminary Design of an Earth-Orbiting Spaceship” is that it was a serious engineering study—about 250 pages long—that predicted nearly all of what has happened in spaceflight since 1946. Weather observations, intelligence gathering, communications, and other things now being done by Earth orbiting vehicles were treated. In addition, there were estimates of what it would take to put people in space. Thus, the stage was set for humanity’s first steps into space.

The Development of U.S. Space Launch Vehicles

As the authors of the 1946 report anticipated, the United States embarked on a vigorous program to develop and then build large liquid-fueled rockets for military reasons. Two military services, the U.S. Army and the U.S. Air Force, were charged with the responsibility for creating the missiles that would carry nuclear weapons. In 1950, the Army reached the decision that the Germans had finished their work at White Sands and that they would be moved to the Redstone Arsenal near Huntsville, Alabama, to initiate the development of new ballistic missiles.

The U.S. Army Rocket Development Program. Wernher von Braun, the leader of the German rocket engineers, arrived in Huntsville on 15 April 1950. It is safe to say that the town has not been the same since. The Army’s attitude toward long-range rockets was that they were an extension of artillery. Thus, in the beginning, the objective was to develop a second-generation V-2 missile. The military requirements, written in early 1951, called for a rocket that had a range of about 200 miles but had a substantially increased payload capability to carry the nuclear weapons then being developed. The rocket was named the Redstone and there were some important differences between the Redstone and the V-2. One was that the payload was designed to separate from the rocket, so that the entire vehicle would not have to reenter the atmosphere. This made it possible to build the rocket from aluminum rather than steel, as was the case for the V-2. The Redstone’s engine had a thrust of 80,000 lb, rather than the 60,000 lb of the V-2. The first Redstone was launched on 20 August 1953, from the new launch site on the east coast of central Florida at Cape Canaveral. The flight was a failure, but subsequent flights proved that the Redstone design was sound. The Chrysler Corporation was given the contract to produce the Redstone in quantity.

Wernher von Braun and his German colleagues never lost their interest in spaceflight. In 1952, von Braun, along with Fred Whipple, the Harvard astronomer, and Joseph Kaplan, a distinguished atmospheric scientist at UCLA, were the leading authors of a series of articles published between 1952 and 1954 in *Collier’s* magazine entitled “Man Will Conquer Space Soon”(5). There were articles about what became the space shuttle, space stations, and journeys to the Moon and the planets. The articles attracted considerable attention and encouraged von Braun and his colleagues to press for an upgrade of the Redstone so that a man-made satellite could be put into Earth orbit. On 15 September 1954, von Braun submitted a proposal to Colonel Toftoy—who by now was chief of the rocket branch at the Redstone Arsenal—to upgrade a Redstone rocket with suitable upper stages so that a small satellite could be put into Earth orbit. This was technically feasible, but the proposal was rejected for political reasons.

The Eisenhower Administration was anxious to keep the effort to create an Earth orbiting satellite as a civilian project. The scientific community had designated 1957 as the first "International Geophysical Year," and orbiting a satellite would be part of the program of research. The satellite would be launched by a Vanguard rocket that would be developed by the Martin Company for this purpose. Even though people from the U.S. Naval Research Laboratory were involved in developing the Vanguard, the program would be managed by a civilian organization.

In spite of this edict, the Redstone was upgraded. The first of these modifications was called the Jupiter A, which was used as a rocket test vehicle. The second, Jupiter C, which had three solid-fueled upper stages, was used in developing heat-resistant materials for ballistic reentry vehicles that would carry nuclear warheads. The first Jupiter C was launched on 20 September 1956 carrying a one-third scale warhead. It reached an altitude of 682 miles and flew a distance of 3,335 miles, a record that stood until the first intercontinental missiles were tested. It was only a short step from the Jupiter C to a rocket that could orbit a satellite. Von Braun secured permission to continue studies to achieve this objective in spite of the fact that his Army superiors were under strict orders not to develop orbital vehicles. Finally, an extended version of the Redstone was used to develop the Jupiter MRBM (medium range ballistic missile) that had a total takeoff thrust of 150,000 lbs.

The Soviets were not bound by the rules imposed on the U.S. Army's group at Huntsville. On 4 October 1957, using a military rocket, the Soviets placed the first man-made object, Sputnik I, into an orbit around Earth. (How this was done is described in an article elsewhere in this *Encyclopedia*.) The Soviet spacecraft created a sensation and caused great consternation in the United States. To add insult to injury, the Soviets launched a much larger satellite, Sputnik II that carried a dog named "Laika" and weighed 1100 lbs, a month later on 3 November. The final embarrassment came on 6 December 1957 when the first Vanguard was launched and then fell back to Earth two seconds later. At this point, the decision was finally made to ask von Braun's group to use the Jupiter C to orbit an American satellite. The Huntsville group recruited William Pickering, the Director of the Jet Propulsion Laboratory at the California Institute of Technology, then funded by the Army, to build a fourth stage for the Jupiter C. Professor James A. Van Allen was recruited to build a payload for Pickering's fourth stage so that if it went into Earth orbit, some scientific results would be obtained. Von Braun's team estimated that they could put a satellite into Earth orbit within sixty days.

The modified Jupiter C, now called the Juno I, blasted off from Cape Canaveral on 31 January 1958 and put Explorer I, the first American satellite into Earth orbit. It was launched 56 days after the Huntsville group was given the job to go ahead. Professor Van Allen's scientific payload was the first to measure the radiation fields above the Earth's atmosphere. This led to the discovery of permanent "belts" of radiation surrounding the Earth, now called the "Van Allen Belts." Juno I was used to launch two more satellites, Explorers 3 and 4, and was then replaced by further upgrades of the Redstone, Juno II, and the Mercury Redstone. Juno II, was used to launch a series of "Pioneer" satellites that explored the newly discovered radiation belts. Finally, the Mercury-Redstone was

used to launch Alan Shepard on a suborbital flight on 5 May 1961 followed by an identical flight by Gus Grissom on 21 July. Both of these flights were carried out after Yuri Gagarin achieved the first orbital flight by a human being on 12 April 1961. (See the article *First Flight of Man in Space* by Klimuk and Vorobyev on Gagarin's flight elsewhere in this *Encyclopedia*.) Thus, the United States was still substantially behind the Soviet Union in spaceflight technology.

The flights of Shepard and Grissom were the last to use the Redstone as the core of the launch vehicle. Thus, the period of playing catchup with the Soviet Union would not be over until launch vehicles more powerful than the Redstone were available.

The U.S. Air Force Rocket Development Program. The U.S. Air Force, unlike the Army, was given a broader mission in rocket development. In addition to short- and intermediate-range missiles, the Air Force would also develop rockets with intercontinental ranges. Another difference between the Army and the Air Force rocket programs was that the leaders of the Air Force program were Americans rather than people from Germany who had been captured at the end of the Second World War. Although a number of members of the German rocket team went to work for contractors who built the rockets, none of them went to work in the Air Force management organization, the Western Development Division in Los Angeles. Walter Dornberger, who had risen to the rank of Major General in the German Army before the end of World War II, went to work at Bell Aircraft, and he was soon joined by Krafft Ehrlicke. Ehrlicke later joined Convair to work on the Air Force Atlas rocket development. Dr. Hans Friedrich also joined Convair to work on the Atlas. Dr. Adolf K. Thiel joined Ramo/Woolridge, later TRW, and Dr. Martin Schilling became a vice president of Raytheon Corporation.

Wernher von Braun and the rocket team that remained with him in Huntsville eventually were transferred to NASA when part of the Redstone Arsenal was turned over to the new civilian space agency in 1960. The new organization was called the George C. Marshall Space Flight Center. There they developed the Saturn V launch vehicle that eventually put humans on the Moon in 1969. (See the article on *U.S. Manned Spaceflight: Mercury to the Shuttle* elsewhere in this *Encyclopedia*.)

The rocket development program that the Air Force eventually adopted grew out of a careful study conducted by the Air Force Strategic Missiles Evaluation Committee. This group was headed by Professor John von Neumann of Princeton University who played a major role in the development of nuclear weapons during World War II. Von Neumann first broached the idea of putting nuclear warheads on top of large rockets, which he called "intercontinental artillery." In 1954, the Missile Evaluation Committee urged the development of relatively small intercontinental ballistic missiles (ICBMs) that were able to carry the advanced "second-generation" nuclear and thermonuclear weapons then being developed. Von Neumann's detailed familiarity with nuclear weapons development influenced the Evaluation Committee to make this judgment. He knew that nuclear explosives much smaller and more efficient than those used at Hiroshima and Nagasaki were being developed, and the rockets, therefore, would be tailored to carry the new weapons. Ironically, this was the reason, among others, that the Soviets gained a significant advantage over the United States

during the first years of orbital space operations. Their nuclear weapon technology was well behind that of the U.S. Accordingly, they had to develop larger and more powerful rockets to carry their heavier nuclear weapons.

To develop the new ICBMs, the Air Force established the Western Development Division at Los Angeles. This move was in accord with the recommendation of the von Neumann Committee to create an organization that would have full authority and responsibility for ICBM development. The Western Development Division would have the technical support of a new organization, the Aerospace Corporation, which was a nonprofit organization created by the transfer to the Air Force of the Space Technology Laboratory of the Ramo-Woolridge Corporation. The first commander of the Western Development Division was a brilliant young Air Force Brigadier General, Bernard A. Schriever. He later achieved four-star rank as the leader of the U.S. Air Force Systems Command. The basic organization described here is still in existence, although the Western Development Division has undergone periodic name changes. Today, it is the Air Force Space and Missiles Division. This organization still receives technical support from the Aerospace Corporation just as it did half a century ago.

The Western Development Division was given the job of developing three missile systems, two ICBMs, the Atlas and the Titan, and one intermediate range ballistic missile (IRBM), the Thor.

The Thor Missile and the Delta Space Launcher. The Thor first stage is a missile that has roughly the same performance characteristics as the Army's Redstone missile. The decision to go ahead with the Thor was controversial because of the apparent duplication of effort. Eventually, the Joint Chiefs of Staff approved the proposal to go ahead with both missiles. In December 1955, the Douglas Aircraft Company was given the contract to develop the missile. The first successful flight of the Thor was carried out on 20 September 1957, and the first operational missiles carrying warheads were deployed in the United Kingdom in 1958, 3 years after the contract was given to Douglas. This remarkably rapid development cycle was due to General Schriever's push for "concurrency." This means essentially that calculated risks are taken in the development process to speed the schedule and, in this case, the adoption of "concurrency" succeeded.

Both the Thor and the Jupiter MRBM had pressurized, regeneratively cooled rocket motors that developed thrusts of about 150,000 lb. Unlike the Redstone, which burned ethanol and liquid oxygen, the Thor burned a kerosene liquid oxygen mixture. It was also designed to be very rugged, and this is the feature that has led to the large number of versions of the Thor, so that it has been called, justifiably, the workhorse of space launchers. It is truly remarkable that space launch vehicles based on the Thor started in the late 1950s with the capability of placing a few hundred pounds in near Earth orbit and now can place a payload of more than 8000 lb. into a geostationary transfer orbit.

The Thor has carried a number of upper stage vehicles. One is the Agena, built originally by the Bell Aerospace Company and later by the Lockheed Corporation. It will be described later. The most important booster stage was the "Delta" solid-fueled rocket built by Aerojet General Corporation that had a thrust level of 8,000 to 10,000 lb depending on the version used. The Delta stage was used so frequently with the Thor that the combination is now called the "Delta."

The second important thrust augmentation for the Thor is the Castor rocket built by Thiokol Corporation. These are powerful solid-fueled rockets that are strapped to the bottom skirt of the Thor rocket. Each unit has a thrust level of 50,000 lbs, and they can be strapped on in numbers between three and nine. The Castor rocket capability adds both power and flexibility to the Delta space launch vehicle system. More than 200 successful Delta rocket launches have been conducted since the first Thor was launched in 1957.

The Delta launch family originated in 1959 when the NASA Goddard Space Flight Center awarded a contract to Douglas Aircraft Company to produce and integrate 12 launch vehicles. Using components from the U.S. Air Force Thor intermediate range ballistic missile (IRBM) program and the U.S. Navy Vanguard launch program, the Delta rocket was available 18 months after the award. On 13 May 1960, the first Delta was launched from Cape Canaveral Air Force Station, Florida, carrying a 178-pound Echo I passive communication satellite. Although the first flight was a failure, the ensuing series of successful launches established Delta as one of the most reliable of all U.S. boosters.

The Delta II Space Launcher. For more than 40 years, the Delta system has consistently demonstrated its robust design, launch flexibility, and value to launch service customers. A second generation, the Delta II launch system, was developed to include multiple configurations to suit the needs of the U.S. Air Force, the National Aeronautics and Space Administration (NASA), and to accommodate the emerging commercial satellite market.

From 1985 through 1987, the space industry was impacted by an unprecedented string of failures of various launch systems, which seriously impeded U.S. space launch capability. In one of several steps to revitalize assured access to space, the U.S. Air Force held a competition for a medium launch vehicle that primarily would launch Global Positioning Satellites (GPS). The contract was awarded to McDonnell Douglas (now Boeing) for its Delta II series vehicles. With a 98.1% mission success rate since its inception in 1989 the Delta II has become the industry workhorse for deploying remote sensing satellites for U.S. government and commercial applications, GPS, commercial satellite systems/constellations, and numerous planetary missions for NASA.

In addition to the demonstrated reliability of the Delta II launch system, Delta vehicles provide incremental performance capability with three, four or nine Castor solid rocket motor (SRM) configurations. These configurations provide a broad range of performance, from 2000 to 4000 pounds to geosynchronous transfer orbit (GTO), using the highly reliable Rocketdyne RS-27A main engine, and two- and three-stage configurations. The latest version, which is designated Delta II Heavy, integrates the solid rocket motors (SRMs) used on Delta III with the Delta II standard upper stage, resulting in approximately a 12% increase in payload lift performance from the standard nine SRM configuration to more than 4700 pounds to GTO (Fig. 3) (6).

The Delta II launch system payload accommodations provide various mechanical interfaces, separation systems, and deployment systems that are designed for compatibility with launch industry standard interfaces. Payload accommodations enable deploying single, dual, or multiple satellites in a single Delta II launch. The multiple manifest, spacecraft dispensers have successfully

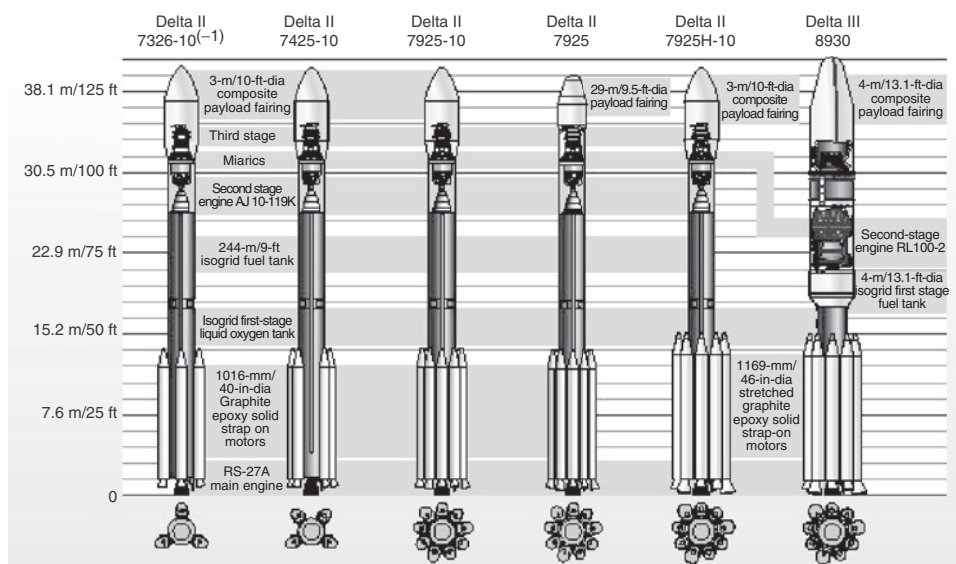


Figure 3. Delta II and III space launch vehicles. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

deployed 55 Iridium spacecraft, five per launch on eleven launches, and 24 Globalstar spacecraft, four per launch on six launches.

The Delta II is launched in the United States from three launch pads, two on the East Coast at Cape Canaveral Air Force Station that has a capacity of 12 launches and one on the West Coast at Vandenberg Air Force Base, California, that has a potential of nine launches per year. Launch sites on both coasts enable the Delta II launch system to launch to virtually any orbit and provides customers with launch schedule assurance.

The demand for Delta II launches increased significantly in the mid- and late 1990s and has continued to maintain a steady backlog of government and commercial customers. In addition to continuing GPS deployment missions, Delta II has been selected for numerous NASA payloads because of NASA's focus on smaller, less expensive spacecraft and its demand for proven reliable launch systems. NASA remains a critical customer for Delta launches, Delta II has assigned and planned missions through 2009.

Building on the success of the Delta II launch system, McDonnell-Douglas developed the two-stage Delta III in the mid- to late 1990s to address the trend toward increasing mass of commercial satellites. The Delta III nearly doubled the payload lift capability of the Delta II launch system by deploying an 8400-pound payload to GTO. Delta III evolved from the highly reliable Delta II by maximizing the use of common components and infrastructure. Both launch systems use the same Rocketdyne RS-27 main engine, first stage LO₂ tank, flight avionics, and launch operations infrastructure. Most notable of Delta III's evolved features are a 13-foot diameter cryogenic second stage, 13-foot diameter composite biconic payload fairing and nine, larger, more powerful SRMs (46-inch dia.). The second stage uses a single RL10B-2 engine fueled by LH₂ and LO₂ that incorporates an extendable nozzle.

The Delta III established itself as an operational launch system in August 2000 by launching a 9460-pound demonstration payload to a planned subsynchronous GTO. The mission was successful; all of the systems and subsystems performed as planned. Intended to be a transition vehicle to the Delta IV, the Delta III has enabled demonstration and flight qualification of several critical components that would be used on the Delta IV.

The Atlas Intercontinental Ballistic Missile (ICBM) and Space Launcher. Concurrently with the development of the Thor, the U.S. Air Force also undertook to develop the first large truly intercontinental missile—the Atlas. This missile included a number of innovative design features. The first and most important of these was that the body of the rocket vehicle was also the wall of the pressurized fuel and oxidizer tanks, and unlike the Thor and Redstone, both of which had conventional braced aluminum air frames, the Atlas skin was manufactured from a thin sheet of stainless steel. The Atlas could only stand on a launch pad if the tanks were pressurized; if they were not, the vehicle would collapse. The pressurized steel tank thus assumes some of the structural burden. This design saves enough weight, so that the Atlas dry weight fraction was the smallest of any large rocket yet built.

The propulsion system of the Atlas was also unique. It consisted of three engines, one mounted on the centerline of the vehicle and the other two fitted on a skirt and placed on either side of the center engine. At launch, all three Rocketdyne MA5 engines would be running, generating a total thrust of 370,000 lb. The skirt, called the “booster section,” could be jettisoned at an appropriate time, and the center engine, called the “sustainer,” would continue to run, generating 60,000 pounds of thrust until the propellant is consumed. This concept is called “stage and a half” because no fuel is carried in the booster section, so that no fuel tank is dropped. The engines burned RP-1 (kerosene) using liquid oxygen as the oxidizer. The engines were gimballed for thrust vector control, as was the Vanguard rocket mentioned earlier.

The Atlas program was initiated in 1951, and the contract to develop the rocket was given to Convair/General Dynamics Corp. The first successful flight of an Atlas ICBM occurred in December 1957, and the first employment as a space vehicle launcher occurred in December 1958 when it was used to put an active communications satellite that weighed about 200 pounds into Earth orbit.

Since these first flights, the Atlas, in combination with a number of different upper stages, has been a very successful space launch vehicle. Perhaps the most important launch of an Atlas booster was John Glenn’s first flight in the Mercury program. The Mercury spacecraft, “Freedom Seven,” was placed into Earth orbit by a modified Atlas D ICBM on 20 February 1962 (Fig. 4). The Atlas has also been used very effectively with Agena, Centaur, Delta, and Burner upper stages. (The technical details of these upper stages will be discussed more thoroughly later.) The Atlas/Agena combination was used for some of the first Mariner Missions (Mariner 4) that returned the first pictures from a Mars flyby in July 1965. The Atlas/Agena was also used to send a Ranger spacecraft to the Moon for a hard landing, and it placed the Lunar Orbiter around the Moon as well. The latter spacecraft was particularly important because the high resolution pictures it obtained were used to select the landing sites for the Apollo landings. The Atlas/Centaur combination is a more capable launch vehicle that was used to put



Figure 4. Launch of John Glenn's Mercury "Freedom Seven" spacecraft by an Atlas space launch vehicle. (photo courtesy NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the first soft landing spacecraft, the Surveyors, on the Moon. These spacecraft were also very important precursors for the Apollo missions because they determined the mechanical properties of the lunar surface.

The most interesting application of the Atlas/Centaur was launching the Pioneer 10 spacecraft on 3 March 1972, as well as its sister ship, Pioneer 11, 13 months later. Pioneer 10 was the first spacecraft to pass beyond the orbit of Mars,

to fly through the asteroid belt, to fly past Jupiter, and finally, it became the first human artifact to leave the solar system. Pioneer 11 repeated this performance a year later, and, in addition, it became the first spacecraft to fly by Saturn.

The Atlas II family was developed in the mid-1980s to address the growing demand for large commercial geosynchronous satellites. The Atlas II family has a 100% success rate and 56 consecutive launches, a reliability record unmatched in the industry. During the past decade, the Atlas II vehicle has been continually stretched and upgraded to improve payload performance. The Atlas IIAS is the most powerful and has the highest lift capability of the Atlas II family. Other configurations include the Atlas IIA and the Atlas II, which was retired in March 1998. Currently, Atlas II vehicles are being flown from both U.S. ranges.

Four solid fuel rockets are used to augment payload performance of the Atlas IIAS. All three MA-5 engines are ignited prior to liftoff. Approximately 180 seconds into first-stage ascent, the two larger MA-5 booster engines are shut down and jettisoned, reducing weight and improving payload performance. The center MA-5 sustainer engine burns for an additional 100 seconds up to main engine cutoff and staging. The Centaur second stage uses two RL-10A-4 engines to place up to 6700 lb (Atlas IIA) and 7950 lbs (Atlas IIAS) into a GTO orbit.

In the early 1990s, General Dynamics (now a part of Lockheed Martin) decided to upgrade the Atlas first-stage propulsion system. The prime modification was replacing the two MA-5 first-stage and single MA-5 sustainer engines with a single Russian NPO Energomash RD-180. Furthermore, Lockheed Martin simplified vehicle construction by drastically reducing the total part count. The reengineered vehicle, known as the Atlas IIAR, has since been renamed the Atlas III. The maiden flight of the Atlas III launch vehicle, the replacement for the Atlas II, occurred on 24 May 2000.

Two versions of the Atlas III are currently available. The Atlas IIIA has a single RL-10A-4-2 engine powering the Centaur upper stage, whereas the Atlas IIIB has two RL-10A-4-2 engines and a stretched Centaur upper stage to increase GTO performance to just under 10,000 lb (Fig. 5) (7).

The Titan ICBM and Space Launch Vehicle. One of the limitations of the Redstone, the Thor, and the Atlas as military missiles was that they could not be kept on "instant alert." This meant that they could not be launched on very short notice because the liquid oxidizer (liquid oxygen) is a liquid only at 210°C below room temperature so that it cannot be stored on the missile itself. Special cryogenic storage facilities had to be built at each of the military launch sites, and upon the order to launch the missile, the liquid oxygen had to be transferred from the storage tank to the missile. Such operations take, at best, something of the order of half an hour, which means that an instant "launch on alert" is not possible.

The Titan missile actually started as a replacement for the Atlas because the fragility of the Atlas was deemed undesirable for deployed military rockets. The Titan was thus designed using the more rugged monocoque technique in which an appropriately braced aluminum "fuselage" took the stresses of the launch. Originally, the first version of this missile, the Titan I, was designed for conventional fuels and would be placed in hardened underground shelters. A successful test was conducted in February 1959, and the Titan I system became

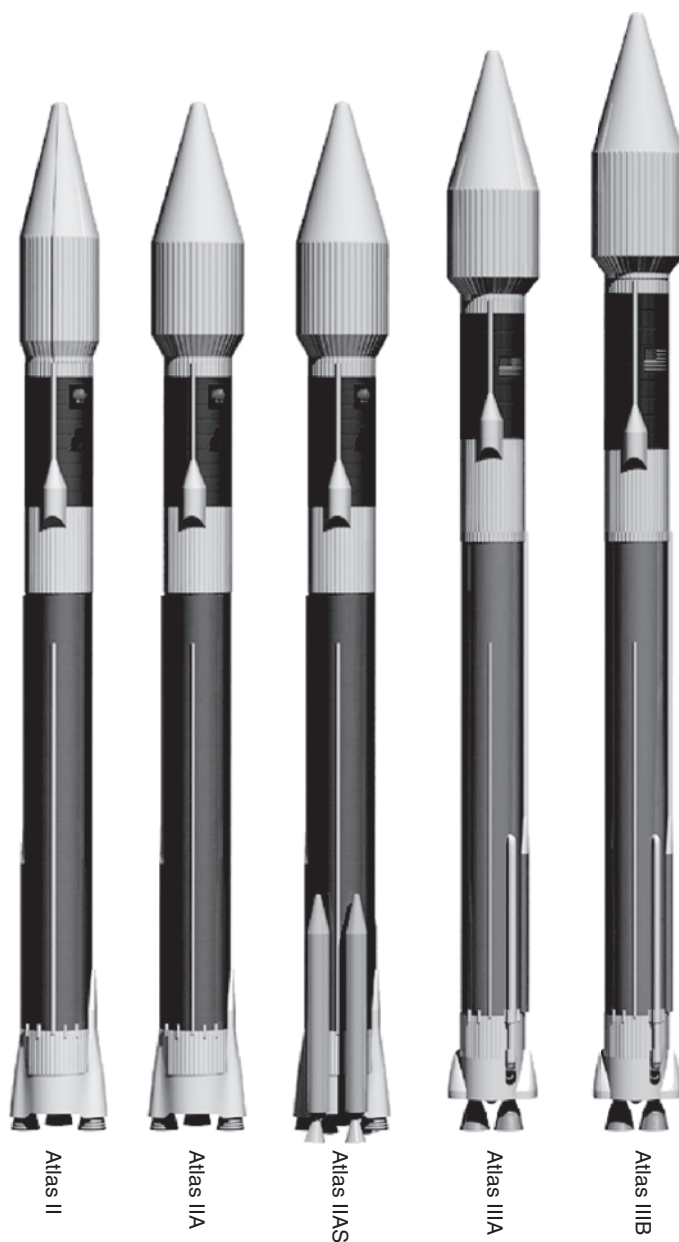


Figure 5. The Atlas space launch vehicle family. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

operational in 1962. In spite of this, the system was awkward to operate, and the Air Force decided to convert the Titan I to a missile that could use storable fuels so that it could be maintained on instant alert. The implementation of this idea led to the Titan II ICBM.

The Titan I ICBM was somewhat more capable in range and payload than the Atlas. The Titan II had to be substantially more capable. The storable fuel

chosen for the Titan II was a mixture of nitrous oxide (N_2O_4) and “Aerozine,” a hydrazine-based fuel. These two substances are liquids at room temperature, and they ignite when they come into contact, that is, they are hypergolic. Thus, with appropriate care, they can be stored on the missile itself. However, the hypergolic fuel mixture does not have the same high specific impulse of the liquid oxygen/hydrocarbon fuels, so that the rocket is not as effective as a launch vehicle. Thus, the Titan II ICBM was designed from the very beginning as a two-stage system, which gave it substantially better performance than the Atlas. The Titan II second-stage motor was also fueled by a hypergolic mixture, so that the entire vehicle could be stored at room temperature.

The first successful test of the Titan II was carried out on 16 March 1962, and the first operational missile was installed in its hardened underground bunker in December 1962. Ultimately, 51 Titan II missiles were deployed at three sites near Wichita, Kansas; Tucson, Arizona; and Little Rock, Arkansas. The Titan II was a formidable part of the U.S. nuclear deterrent force; it carried a multimegaton warhead that could reach targets 8000 to 10,000 miles away. In September 1980, a technician at one of the Titan II missile sites in Arkansas dropped a wrench down the silo and it punctured the fuel and oxidizer tanks as it fell. Eventually, the two liquids mixed, ignited, and the resulting conflagration destroyed the missile and silo. Fortunately, the nuclear warhead, which had a hardened design, survived the accident unharmed. Partly as a result, all of the deployed Titan II ICBMs were decommissioned by 1987.

Like the Thor and the Atlas, the Titan II ICBM was also converted into a very flexible and capable space launch system. The first of these, fielded in 1964, was called the Titan IIIA; it consisted of the Titan II two-stage ICBM plus a third stage called “Transtage” that could put payloads of more than 3000 pounds into low Earth orbit. This was followed quickly by the Titan IIIB/Agena in 1966 that could put 8500-pound payloads into near Earth orbit. This was accomplished by increasing the thrust of the first stage from 430,000 to 463,000 pounds and using the somewhat more capable Agena rather than the Transtage. All of these modifications were carried out by the Martin Company.

The Titan II space launch vehicle is a two-stage liquid-fueled booster designed to provide small-to-medium weight class capability. It can lift approximately 4200 pounds into a polar, low Earth circular orbit. Titan IIs were also flown in NASA’s Gemini manned space program in the mid-1960s. Deactivated Titan II missiles are in storage at Davis-Monthan Air Force Base in Tucson, Arizona. Lockheed Martin was awarded a contract in January 1986 to refurbish, integrate, and launch 14 Titan II ICBMs for government space launch requirements.

Tasks involved in converting the Titan II ICBMs into space launch vehicles included modifying the forward structure of the second stage to accommodate a payload; manufacturing a new 10-ft diameter payload fairing with variable lengths plus payload adapters; refurbishing the Titan’s liquid rocket engines; upgrading the inertial guidance system; developing command, destruct, and telemetry systems; modifying Vandenberg Air Force Base Space Launch Complex 4 West to conduct the launches; and performing payload integration. Six Titan II Space Launch Vehicles have been launched from Vandenberg Air Force Base, California, since 5 September 1988.

The major modification of the Titan III could be made because of the sturdy monocoque construction of the original Titan II missile. This was the addition of two very large solid-fueled strap-on rockets on opposite sides of the Titan II core vehicle. These rockets have a thrust of just over 1,000,000 pounds each, which raises the takeoff thrust of the entire vehicle in the launch configuration to about 3,000,000 pounds. This configuration of the launch vehicle is called Titan IIIC, and with a Transtage, it can place about 30,000 pounds in low Earth orbit. It can place a payload of about 3600 pounds in a geosynchronous orbit by using the appropriate upper stages. Another important variation is the Titan IIID, which has no upper stages but can put a 13,000-pound payload into near Earth polar orbits. This is the launch vehicle that, when launched from the West Coast at Vandenberg Air Force Base, puts most of the U.S. highly capable reconnaissance satellites in orbit. The Titan IIID was declared operational in 1971.

In 1977, the Titan IIIE/Centaur was fielded. This is the most capable expendable space launch vehicle in the current American inventory. The addition of the Centaur upper stage makes the difference (Note: The upper stage will be described in more detail later.) The Titan IIIE/Centaur has launched a whole galaxy of spacecraft to explore everything from the outer planets using exquisitely designed cameras to putting very sophisticated payloads into the atmosphere of Jupiter. This would include Helios, Galileo Jupiter with the probe, the two Voyagers, the two Vikings, and a substantial number of others. A related version of the Titan IIIE is the Titan 34D which, instead of carrying the Centaur, carries a solid-fueled upper stage called the "inertial upper stage" or IUS. This stage was developed by the Air Force to put large military and intelligence gathering satellites into polar orbits.

Titan IV consists of two solid-propellant, stage-zero motors, a liquid propellant two-stage core and a 16.7-ft diameter payload fairing. Upgraded three-segment solid rocket motors increase the vehicle's payload capability by approximately 25% (Fig. 6) (7). In 1985, the U.S. Air Force selected the Martin Marietta Astronautics division (now Lockheed Martin) in Denver to build and launch 10 Titan IVs. In 1986, the contract was increased to 23 vehicles, and, in November 1989, the contract was extended to 41. Titan IV has been used exclusively to launch U.S. government satellite missions. It provides primary access to space for critical national security and civil payloads.

The operating success rate of Titan launch systems is better than 95%. It can place 47,800 lb into low Earth orbit or more than 12,700 lb into geosynchronous orbit.

Titan IV is launched from Launch Complex 40 at Cape Canaveral Air Force Station, Florida, and from Space Launch Complex 4E at Vandenberg Air Force Base, California. The first Titan IV B was successfully flown from Cape Canaveral Air Force Station on 23 February 1997. This configuration improves reliability and operability and increases lift capability by 25%. Advancements also include improved electronics and guidance. The Titan IV B has standardized vehicle interfaces that increase the efficiency of vehicle processing. Additionally, the more efficient programmable aerospace ground equipment is used to monitor and control vehicle countdown and launch.

Upper Stages. In describing the evolution of military missiles to become space launchers, we mentioned a number of upper stages used in combination

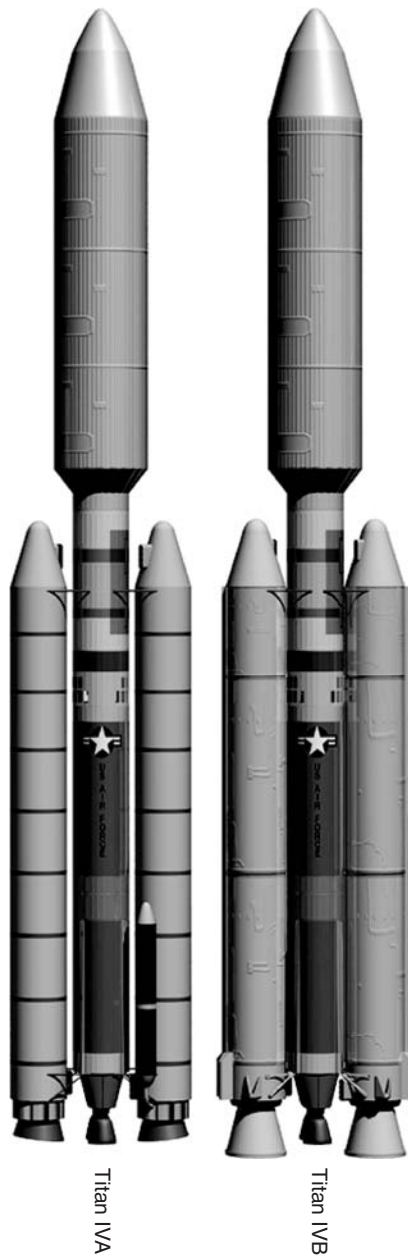


Figure 6. The Titan IV space launcher. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

with the military rockets to place spacecraft into Earth orbit and beyond. In this section, we will provide brief descriptions of the most important.

The Centaur, developed and built by Convair/General Dynamics, is the most capable of these vehicles. The Centaur uses liquid oxygen to burn hydrogen fuel. This mixture is the most potent of all chemical rocket fuel combinations

because it provides the highest specific impulse to the rocket. Thus, rockets using this combination have the highest propulsive efficiency. The Centaur can be shut down and restarted in space, which is most important because it permits executing complex space maneuvers. However, because the fuels are cryogenic, these maneuvers have to be carried out shortly after launch, so that the fuel and oxidizer are not lost by evaporation. The Centaur has two Pratt and Whitney RL-10 engines that develop a total thrust of 30,000 pounds. The first Centaur was flown in 1962, and since then, Centaur upper stages have been used on top of Atlas and Titan boosters.

The Agena is the second important liquid-fueled upper stage. The Agena also uses a hypergolic fuel/oxidizer mixture of (N_2O_4) and unsymmetrical dimethyl hydrazine (UDMH) to provide a thrust of about 15,000 pounds. The Agena also has shutdown and restart capability like the Centaur. The difference is that the fuel/oxidizer mixture can be stored at ambient temperature, so that the Agena vehicle can stay in orbit in a dormant condition much longer than the Centaur. The Agena has been used on Thor, Atlas, and Titan rockets and has proven to be the most ubiquitous of the high-performance liquid upper stage vehicles. The Agena was originally designed and built by the Bell Aerospace Company for the Atlas and was later taken over by the Lockheed Missiles and Space Company (now called Lockheed Martin Co.). It was flown for the first time in 1960 on an Atlas booster.

Solid-fueled rockets have been used extensively as upper stages for space launch vehicles ever since the first orbital flights. The Juno I launcher, which placed Explorer I in Earth orbit in January 1958, had three solid-fueled upper stages, including a small solid booster on the Explorer I satellite itself. Solid-fueled rockets have the advantage that they are safely storable and also extremely versatile. They have ranged from the small rockets already mentioned that might develop tens of pounds of thrust to the huge strap-on rockets of the Titan IIIC that develop thrust levels of more than a million pounds. The primary disadvantage of solid-fueled rocket boosters is that, except in certain cases, such as the two-grain systems used in antimissile applications, they cannot be turned off once they are lit. Another disadvantage is that failure of solid-fueled rocket is usually catastrophic—which was unfortunately graphically illustrated in 1986 by the failures of a solid rocket booster in January that caused the loss of “Challenger” and then again in April that caused the loss of a Titan 34D flight.

There are too many solid-fueled upper stages to list here. Probably the most important and capable solid-fueled upper stage is the inertial upper stage (IUS) that was built by the Air Force for use with the Space Shuttle and the Titan 34D system.

The Evolved Expendable Launch Vehicle Program (EELV)¹

The EELV Concept. The Evolved Expendable Launch Vehicle program, born of studies conducted in the late 1980s and 1990s, represents a commitment to

¹Excerpts from input by James Simpson, The Boeing Company

reducing significantly the cost of access to space. An industry/government partnership has developed two competing EELV systems to meet space transportation needs during the next 20 years (8).

Conceived as a "system of systems" to improve operability and reduce recurring and infrastructural costs, EELV is using streamlined manufacturing and improved mission assurance processes. Its facilities and operations are designed to lower costs.

Requirements call for a 25–50% reduction in recurring operational cost compared to current systems and for improving system reliability and availability. The U.S. Air Force interest is that EELV will replace the Titan, Atlas, and Delta vehicles and their launch infrastructures supported by DOD. The program implements DOD acquisition excellence goals by streamlining the government's role and replacing its oversight of contractors with less intrusive "insight." The objective is to enhance U.S. launch industry competitiveness in the international market by reducing costs across the entire system.

In October 1998, the government awarded \$500 million contracts each to Boeing and Lockheed Martin. Development costs are shared between the contractors and the government, resulting in a national, dual-use launch service. The government program office has virtually unlimited access to all but some highly sensitive and proprietary cost and pricing data. The Air Force simultaneously awarded initial launch service contracts to both firms.

The strategy enabled two further benefits: competition and assured access to space. Having two competitors throughout the life cycle of the program is key to achieving price competitive procurement. Two providers using a standard payload interface maintain payload interchangeability between Delta IV and Atlas V and enhance assured access to space.

Each delivery order for a launch service has a standard 24-month period for performance. Individual launch service plans, however, are highly flexible and can be tailored to spacecraft customer needs.

EELV will support U.S. military intelligence, civil, and commercial mission requirements using contractor-provided commercial launch services. The two are the Boeing Delta IV and Lockheed Martin Atlas V, both designed to meet the full range of government launch requirements.

The EELV program has three key performance parameters: specific payload mass-to-orbit requirements; vehicle design reliability of 0.98 (threshold) at a 50% confidence level; and standardization, including standard payload interface for each class of vehicle and standard launch pads that can accommodate all configurations in an EELV family (9).

Delta IV. Delta IV was developed under a U.S. Air Force EELV contract. The Delta IV launch vehicle uses a new liquid oxygen/liquid hydrogen 16.7-foot diameter common booster core (CBC) powered by the new Boeing-Rocketdyne RS-68 main engine. This is the first large liquid-fueled engine to be developed in the United States since the SSME (Space Shuttle main engine). The RS-68 is a gas generator liquid oxygen/hydrogen booster engine. The bell-nozzle RS-68 develops 650,000 pounds of sea level thrust and uses a simple design approach that has drastically reduced the total part count compared to engines of equivalent size

or performance. The vehicle's cryogenic upper stage, which uses the Pratt & Whitney RL10B-2 engine, is substantially similar to that flown on the Delta III (6,9,10).

There are several variants of the Delta IV launch vehicle. The Delta IV-M (medium) is a single-core variant that combines the CBC with a version of the Delta III liquid oxygen/hydrogen second stage and a stretched 4-meter fairing that provides 9200 pounds to GTO. The Delta IV-M + variants augment the single core with two or four solid rocket strap-on Alliant Techsystems graphite epoxy motors (GEMs) and provide two variations of upper stages and payload accommodations, the 13-foot Delta III derivative and the 16.7-foot version that has greater fuel capacity and greater payload volume. The M + variants enable payload deployment of 14,700 pounds to GTO. The Delta IV vehicle family will have a GTO lift capability of up to 29,500 pounds and is available in three major variants. The largest variant, the Delta IV-Heavy, combines three CBCs with the 16.7-foot upper stage. The payload accommodations include either a 16.7-foot isogrid aluminum fairing based on the existing Titan IV or a newly developed composite fairing based on the Delta II and Delta III designs. The 13-foot fairing is the existing Delta III composite fairing lengthened by 3 ft (Fig. 7) (6).

Improvements include the new CBC, the newly developed and simplified main cryogenic engine, the focused factory facility, and simplified launch-processing operations.

Parts for the medium-plus and heavy CBCs are, respectively, 88% and 93% common relative to the medium CBC. All are manufactured using a common factory production list. CBC innovations include friction stir-welded tanks, spun-formed domes, and the use of composite structures. The RS-68 has reduced operating pressure, 80% lower part count, 95% less labor, uses cast versus welded parts, and has no special coatings. However, more than 85% of the upper stage part count is a Delta III heritage, and much of the avionics are from Delta II and III.

Full integration, assembly, and checkout testing take place before each vehicle leaves the factory. Delta IV's horizontal booster processing flow and vehicle stage mating in the horizontal integration facility allow parallel integration, reduced hazardous lifting operations, and decreased pad time. Total vehicle time at the launch base is less than 1 month and only 8–11 days on the pad. Delta IV features launch sites on both the East Coast (Cape Canaveral Air Force Station, Florida) and West Coast (Vandenberg Air Force Base, California). Each pad can launch all configurations, and launch pads are virtually standard between the Cape Canaveral SLC-37 and Vandenberg SLC-6 launch sites.

Atlas V. The Atlas V vehicle family, developed under the U.S. Air Force EELV contract, builds on the improvements made for the Atlas III. The Atlas V family of vehicles incorporates a reinforced first-stage structure, as well as increased propellant load in the first stage, called the common core booster (CCB) powered by the Russian RD-180 engine. The RD-180 is produced by RD AM-ROSS, a joint venture between Pratt & Whitney and Russian's NPO Energomash. The engine develops 860,000 pounds of thrust at sea level, uses liquid oxygen/RP-1 propellants, and is the only high-thrust, staged-combustion engine in production. It has

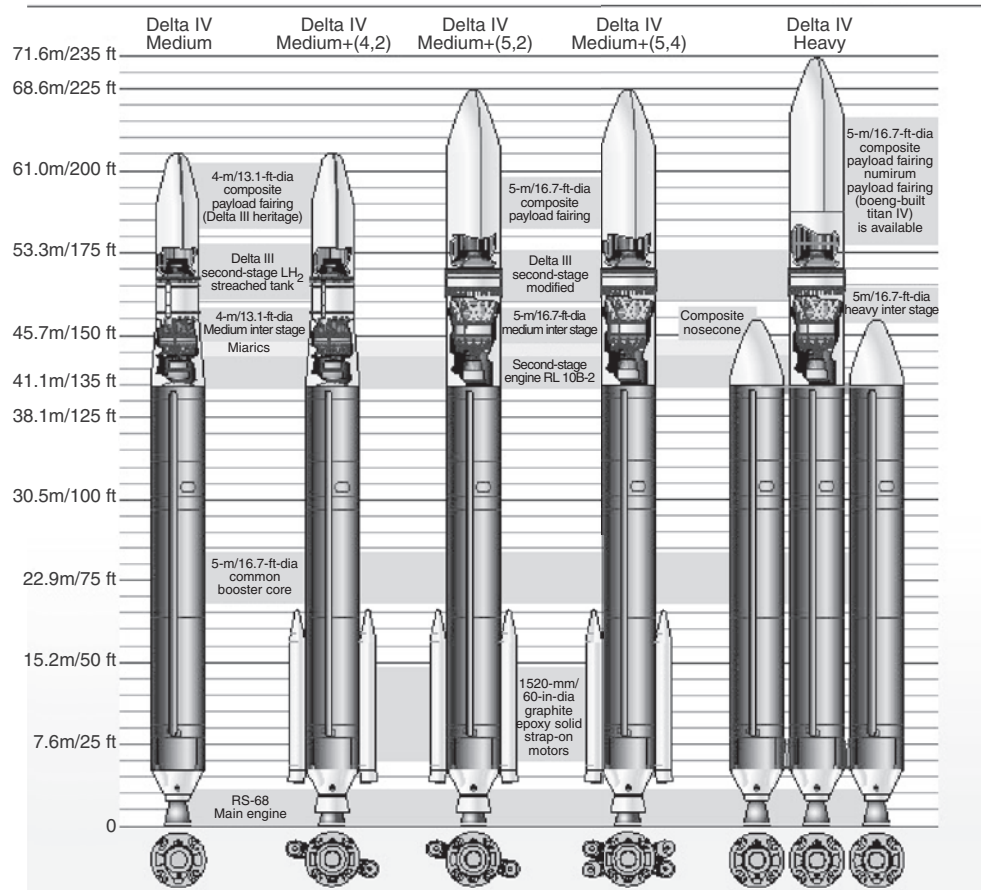
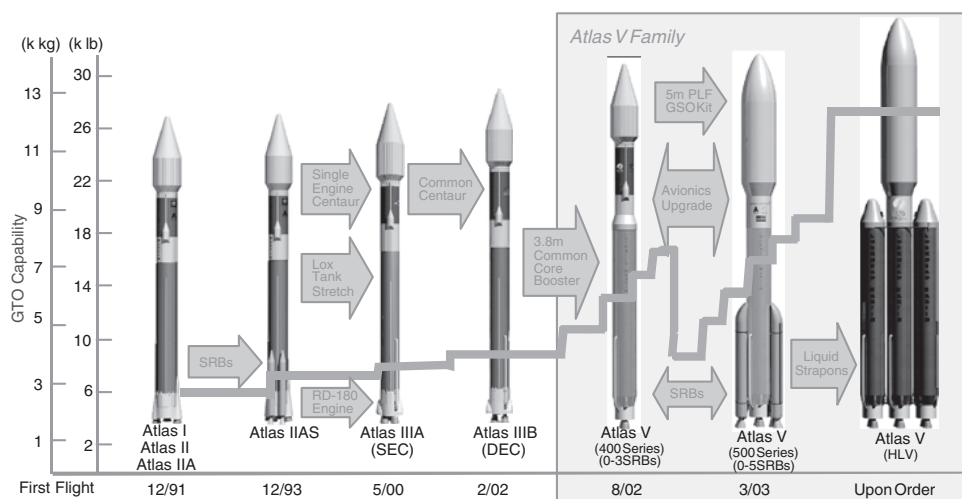


Figure 7. Evolved Delta. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

been tested extensively and was flight proven on the first Atlas IIIA mission in May 2000 (Fig. 8) (7).

Lockheed Martin has proposed variants of the Atlas V that incorporate different arrangements of solid strap-on boosters to increase the payload performance of the single common core variant up to 18,000 lb to GTO. To differentiate the various Atlas V configurations, Lockheed Martin devised a secondary numbering system. The first number identifies the fairing diameter in meters (3,4, or 5-meter fairing). The second number identifies the number of solid strap-on boosters (0 through 5). The final number identifies the number of second stage RL-10 engines (either 1 or 2). As an example, an Atlas 5 532 has a 5-meter fairing, three solid strap-on boosters, and two second-stage RL-10 engines. A single engine RL-10 second stage is used for high altitude (MEO and GTO) missions, whereas the two-engine variant is used for LEO missions.

Atlas V's several configurations have the flexibility to meet varied performance requirements for missions from LEO to GTO. Options include the addition



WSC October, 2002

1

Figure 8. Evolved Atlas. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

of one to five Gencorp Aerojet strap-on solid rocket motors for intermediate lift capability or the use of three CCBs for heavy payloads. The Atlas 400 series has a 13-foot payload fairing and a single CCB; the 500 series has a composite 16.7-foot payload fairing, a single CCB, and up to five Aerojet solid rocket boosters; and the heavy launcher has three CCBs and a composite 18-foot payload fairing. All three series use a common Centaur upper stage with Pratt & Whitney RL10A-4-2 engines(s).

These modifications, combined with the stretched Atlas IIIB Centaur upper stage, allow the Atlas V to place more than 10,000 lb in a geosynchronous transfer orbit with the single CCB. The heaviest variant of the family currently planned for production will incorporate an arrangement of five solid strap-on boosters to increase the payload performance of the single common core variant up to 18,000 pounds to GTO. Lockheed Martin has designed a three common booster variant capable of placing more than 13,000 lb directly into a geostationary orbit that is currently not being marketed commercially (8,10).

Atlas V, which uses the same Centaur upper stage as the Atlas IIIB, can be configured with either one or two RL10A-4-2 engines. A hydrazine attitude control system provides precise in-orbit maneuvering. The 18-foot payload fairing is a new design derived from the Ariane V fairing manufactured by Contraves Space of Switzerland. It will be offered in two lengths, one optimized for communications satellites and the other for accommodating large-volume spacecraft missions. The 13-foot payload fairing is the same one used on Atlas II and III. Among Atlas V's innovations is the RD-180 engine capability for continuous throttle between 47 and 100% of nominal thrust, which allows substantial control over launch vehicle and payload environments. Others include reduced manufacturing cycle time and simplified launch processing. Atlas V also includes

the Air Force EELV standard payload interface that allows payload interchangeability with Delta IV (8).

Atlas V incorporates efficient launch site processing, including use of an off-pad vertical integration facility (VIF) for the vehicle and parallel processing of the encapsulated payload in separate installations. Launch site processing has been reduced from 28–38 days for Atlas II to just 18–26 days. The encapsulated payload will be transported to the VIF and mated to the launch vehicle. After combined systems testing, the fully integrated Atlas V/encapsulated payload will be transported to the nearby “clean launch pad.” All vehicle configurations use common processing procedures and can be launched from the same clean pad. On-pad time has been reduced to less than 1 day.

Current Status. The Lockheed Martin-built Atlas 5 401 (single CCB, no solids, and single engine Centaur) booster launched flawlessly on 21 August 2002 deploying the Hot Bird 6, a communications satellite for Eutelsat.

The 12.5-foot diameter rocket was the largest to launch from Cape Canaveral since the Saturn 5 sent Apollo astronauts to the Moon. This flight gives the Atlas family a string of 61 consecutive successful launches during 9 years using the Atlas II, Atlas III, and Atlas V vehicle configurations.

On 20 November 2002 the Delta IV-M + 4,2 carried the W5 communications satellite for Eutelsat to a precise geostationary transfer orbit.

The Department of Defense and related agencies (e.g., NASA), as well as the commercial sector, will be major customers for both vehicles. The Pentagon has scheduled 29 launches aboard the rockets to date.

Conclusions

It is truly remarkable that from Goddard's first liquid-fueled rocket flight to the landing on the Moon was a mere 43 years—well within the lifetimes of many people living today. The spurs for this achievement were clearly the Second World War (1939–1945) and later, the Cold War (1948–1991). In both cases, the technology of rocket propulsion was considered critical by all concerned to prevailing in the conflicts. However, this is only half the story. The other half is that a group of unusually talented and motivated people from many nations contributed to the successes that we have described. The principal technical conclusion that can be drawn from what has been said here is that the technology of chemically fueled rockets is mature. For the past 40 plus years, since the development of the ICBMs in the early 1960s, no new propulsion technology has been developed and applied. What has happened is that the proliferation of liquid- and solid-fueled chemical rockets has made it possible to develop a very large number of launch vehicle combinations tailored to meet a great many different requirements for expendable space launchers. This, of course, is what is meant by the term “evolution” in the title of this article. For the foreseeable future, this evolution will continue and will be made possible by advances in guidance and control systems, more accurate timing and navigation, and other auxiliary technologies. It is a tribute to the designers of the early ICBM rockets that their products are still in our front line expendable space launchers.

What of the future? We believe that there are now signs on the horizon that new technologies for space launch systems will be required. We are on the threshold of initiating human exploration of the solar system. The International Space Station will be the staging base for this new phase of space exploration. We believe that both electric propulsion systems and nuclear rockets will be assembled at the space station and will eventually take people and equipment on journeys around the solar system. Hopefully, the designers and builders of these new propulsion systems will display the same skill and virtuosity as the people who created the space launchers described in this article.

BIBLIOGRAPHY

1. Goddard, R. A method of reaching extreme altitudes. *Smithsonian Miscellaneous Collections*, 71 (2) (1919).
2. Oberth, H. *Die Rakete zu den Planetenraumen* (The Rocket Into Planetary Space). Oldenburg, Munich, 1923, reprinted by Uni-Verlag, Nürnberg, 1960 and *Wege Zur Raumschiffahrt* (Ways Toward Space Travel), 1929, reprinted by Kriterion, Bucharest, 1974.
3. Ordway, F. III and Sharpe, M.R. *The Rocket Team*. MIT Press, Cambridge, 1979.
4. Clauser, F. and D. Griggs, L. Ridenour, et al. Preliminary Design of an Experimental Earth-Orbiting Spaceship", Report No. 5M-11827 (Contract W33-038 ac 14105, Douglas Aircraft Company, Inc. May 2, 1946.
5. von Braun, W., F.L. Whipple, J. Kaplan, H. Haber, W. Leyy and C. Ryan. Man will conquer space soon. *Collier's Magazine*, a series of articles starting on March 22, 1952 and ending April 30, 1954.
6. Delta and SeaLaunch Technical Summary, The Boeing Company, Huntington Beach, CA, April 2002.
7. Atlas and Titan Data, Lockheed-Martin Website, Astronautics Operations, Denver CO, September, 2002.
8. Knauf, J.M., L.R. Drakee and P.L. Portonova. EELV evolving toward affordability. *Aerospace America*, 38-42 (March 2002).
9. Sietzen, F. Air Force switches to dual production of EELV. *SpaceCast News Service*, 8 November 1997, Washington D.C.
10. Atlas 5-Summary. Space and Tech Database Expendable LVs, www.spaceandd-tech.com/spacedata/elvs/delta4, copyright 2001-Andrews Space & Technology.
11. Delta IV-Summary. Space and Tech Database Expendable LVs, www.spaceandd-tech.com/spacedata/elvs/delta4, copyright 2001-Andrews Space & Technology.
12. Delta IV Payload Planners Guide. The Boeing Company, April 2002.
13. Prandini, E. Arianespace under pressure. *Interavia Business Technol.* 57: 665, 54 (3): (2002).

MICHAEL I. YARYMOVYCH
Boeing Space and Communications (Retired)
Seal Beach, California

HANS MARK
Austin, Texas

EXPLORATION OF MARS BY THE USSR

Introduction

Mars research began long before the Space Age. Ground-based astronomical observations using photometry (I.K. Koval', N.P. Barbashev, et al.), polarimetry (A.V. Morozhenko et al.), and infrared spectrometry (V.I. Moroz) were performed in the Soviet Union throughout the 1950s and 1960s. G.A. Tikhov attempted to find evidence of life on Mars using spectroscopic techniques. Theoretical models for the internal structure of the planet began to be developed (V.N. Zharkov et al.). An overview of this period may be found in Ref. 1.

Early Soviet planetary spacecraft were developed by Special Design Bureau 1 (OKB-1, now NPO Energia) under the leadership of Chief Designer S.P. Korolev (1907–1966). B.E. Chertok has brilliantly described this period of history in his books (2–4). One of the primary motivations for the development of long-term Martian research programs was a desire to find life on the planet. It was generally understood that this project had to begin elsewhere—learning more about the planet Mars itself. Nevertheless, early Soviet spacecraft carried a spectrophotometer for observations of the so-called Sinton bands, which were, at that time, presumed to indicate the presence of organic material on Mars.

Eight attempts were made to launch a spacecraft to Mars, starting in 1962. Only the fourth spacecraft, the Mars-1 (1962) executed a flyby, but communications were lost prior to the flyby. Jet Propulsion Laboratory made two attempts (1964) during this period, one of which was successful: Mariner-4 performed a Mars flyby and transmitted an image of the planet.

In 1965, S.P. Korolev transferred all unmanned interplanetary space flight projects to another facility, the Design Bureau and Plant named after S.A. Lavochkin. The Design Bureau ("KB") was managed by Chief Designer G.N. Babakin (1914–1971). Later (1974) this facility was re-named to Nauchno-Proizvodstvennoe Obyedinenie imeni S.A. Lavochkina (NPO Lavochkin or NPOL), in translation Research-Industrial Association S.A. Lavochkin. For simplicity we will use the designation "NPO Lavochkin" for all periods covered below.

Many successful missions were carried out by NPO Lavochkin to the Moon, Venus, and eventually to Halley's Comet, but, the USSR's development of Mars projects was unsuccessful. Between 1969 and 1973, six Soviet spacecraft were launched, two in 1971 and four in 1973; however, virtually all of the new knowledge about Mars obtained by the early 1980s came from the NASA Mariner 9 and Viking missions. The Soviet contribution to research concerning the planet itself turned out to be quite meager, except for various issues relating to the interaction of the planet and its plasma sphere with the solar wind, where the Soviet Mars-3, Mars-5, and Phobos-2 spacecraft provided the basic results. It should also be noted that the Soviet Union was responsible for the first successful landing on Mars (Mars-2) and the first direct measurements performed in the Martian atmosphere (Mars-6). A NASA-published history of the NPO Lavochkin Mars efforts may be found in Ref. 5. The first papers discussing the scientific results of

the 1973 Soviet missions were published in a special volume of the journal *Kosmicheskie Issledovaniya* [*Space Research*] (6). A description of the major scientific research results related to the Martian atmosphere and surface obtained during both Soviet and American missions of that era may be found in Ref. 7.

Mars-2 and Mars-3 Spacecraft

The Mars-2 and Mars-3 unmanned interplanetary spacecraft were launched on 19 and 28 May 1971. Each spacecraft included an orbiter and a lander (Figs. 1 and 2). Upon arrival at Mars (on 27 November and 2 December 1971, respectively), the landers separated from the Mars-2 and Mars-3 spacecraft and reached the surface of the planet, thereby becoming the first spacecraft to attempt a landing on the surface of Mars. Each 1000-kg lander carried a pennant with the seal of the Soviet Union, a camera, and scientific instrumentation for studying soil samples. These were the first items made on Earth to land on Mars. The Mars-2 lander broke up at coordinates 44.2°S 313.2°W because of inaccurate targeting, which led to an error in the entry angle and caused the lander to impact on the surface before the parachute opened. The Mars-3 lander successfully landed on the surface at coordinates 45°S 158°W. The Mars-2 and Mars-3 orbiters (each weighing 2265 kg) were inserted into Mars orbit and performed a research program on the composition of the Martian atmosphere, surface photometry and IR radiometry, the magnetic field of Mars, and the interaction between the solar wind and the Martian plasma environment for more than 8 months. Both orbiters had cameras but no scientifically important images were transmitted. A global dust storm prevented imaging during first months after orbit insertion, and cameras failed by the time when dust dissipated. The Mars-2 and Mars-3 orbiters ceased operations in August 1972.

Mars-4 and Mars-5 Spacecraft

The Mars-4 and Mars-5 spacecraft for planetary studies of Mars from Mars orbit were launched on 21 and 25 July 1973 (Fig. 3). Mars-4 reached Mars on 10

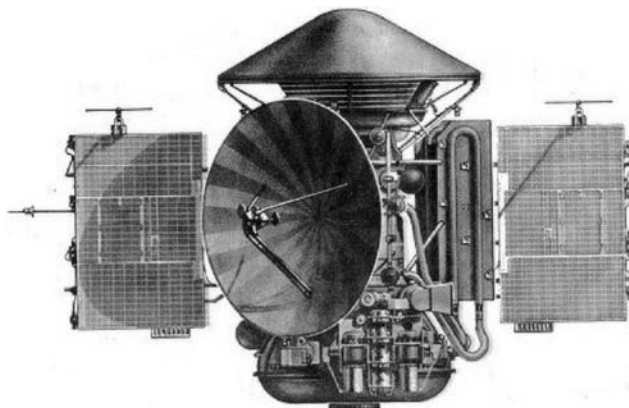


Figure 1. General view of Mars-2 and Mars-3 spacecraft. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

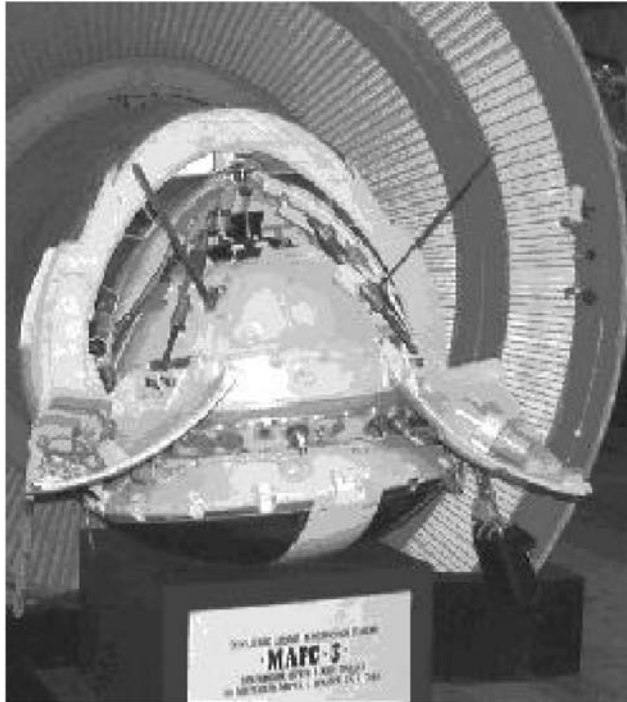


Figure 2. Mars-2 and Mars-3 landers. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

February 1974 and passed within 2200 km of the planet's surface without entering Mars orbit; however, it did transmit photographs of the planet to Earth. The Mars-5 spacecraft entered Mars orbit on 12 February 1974 (the orbiter had a mass of 2200 kg). The spacecraft continued operating once it reached Mars orbit; data were obtained on the atmosphere and surface temperature patterns, and approximately 120 surface photographs and panoramas of the Martian Southern Hemisphere were transmitted to Earth.

Mars-6 and Mars-7 Spacecraft

The Mars-6 and Mars-7 Mars spacecraft were launched on 5 and 9 August 1973. (Figs. 4 and 5). Mars-6 reached the planet on 12 March 1974. The descent module separated from the orbiter, landed on the surface of Mars, and performed direct measurements of the Martian atmosphere during its descent. Mars-7 reached the planet on 9 March 1974. The descent module separated from the spacecraft and performed a flyby 1300 km above the Martian surface.

After 1973, there were no USSR flights to Mars for many years. At that time, it would have been difficult to mount a standard mission (e.g., with a satellite and several lander modules) that could have obtained new scientific results at a level higher than that already achieved by the United States. There was a desire to undertake a project that would be more significant on a

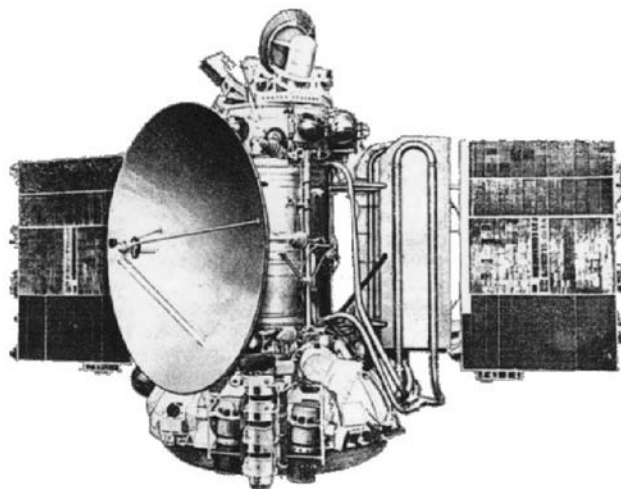


Figure 3. Mars-4 and Mars-5 orbiter. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

fundamental level. The project that was eventually selected involved the return of a sample of Martian material to Earth. NPO Lavochkin worked on this mission for several years, but the project turned out to be too risky and was eventually cancelled (5).

During this period, the USSR successfully continued its flights to Venus. A new direction for research was pioneered by R.Z. Sagdeev. These were research projects related to two solar-system bodies, Phobos and Halley's Comet.

Phobos-1 and Phobos-2 Spacecraft

The primary purpose of the Phobos spacecraft was to study not Mars itself, but Phobos (8). However, this mission was the next major post-Viking step in Mars research and marked the first mapping of the planet in the thermal infrared (the spatial resolution was approximately 1 km at a wavelength of approximately 10 μm). Approximately 40,000 near-infrared (1–3 μm) images were obtained using a mapping spectrometer. At that time, no instruments of this type had yet been used, even in remote sensing observations of Earth. This mission also marked the first spectroscopic remote sensing of the Martian atmosphere via measurements of the solar spectrum at various altitudes above the limb in several wavelength bands; these data were used to determine the vertical water vapor and aerosol distributions. The first measurements of the dissipative flux in the upper atmosphere of Mars were made. Several interesting new results were also obtained for Phobos. Precise mass and density values were determined, and brightness was measured as a function of wavelength. The brightness as a function of wavelength did not conform to previously held views regarding the composition of Phobos. A large amount of unique data was obtained on the plasma environment of Mars and its interaction with the solar wind (9). Papers describing the scientific results of the Phobos mission were published in the journals

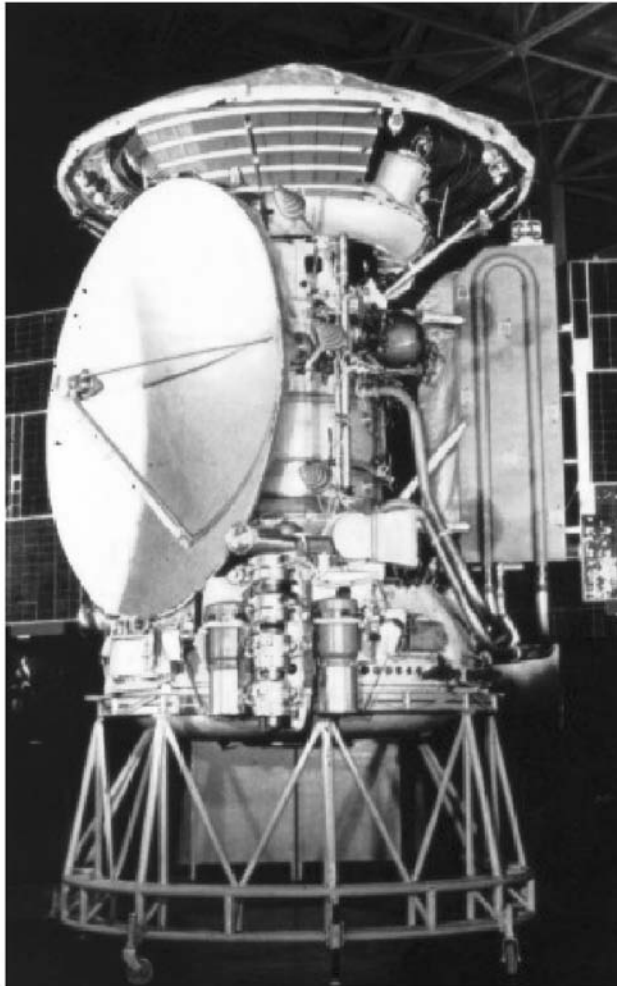


Figure 4. Mars-6/Mars-7 spacecraft. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Nature (10), *Geophysical Research Letters* (11), and *Planetary and Space Science* (12). Despite the fact that Phobos-1,2 was a short-term mission (transmitted data from orbit for approximately 2 months), the resulting scientific data exceeded the scope and quality of that obtained in all of the other Soviet Mars expeditions taken together.

The Phobos-1 and Phobos-2 new-generation spacecraft (Table 1) were launched on 7 and 12 July 1988, respectively (13) (Figs. 6 and 7). These spacecraft performed research on Phobos, Mars, and outer space (Fig. 8). Thirteen countries, plus the European Space Agency (ESA), were involved in developing the scientific instrumentation (Table 2).

Snyder and Moroz (14) provide information on the scientific instrumentation for all Mars missions (including Soviet), as well as a review of the major scientific results.

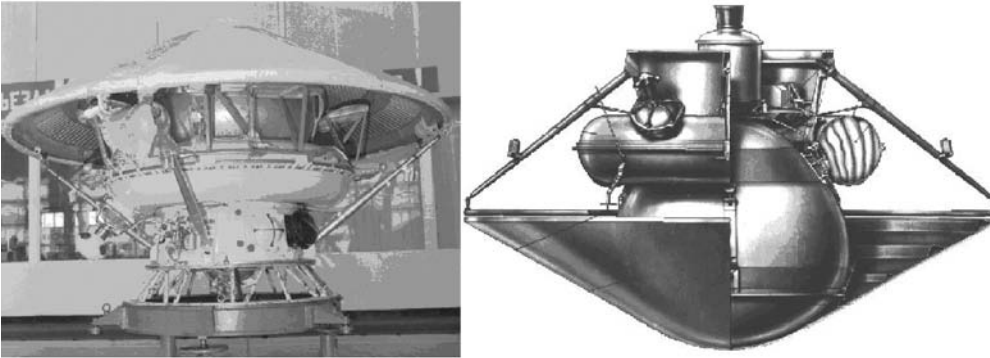


Figure 5. Mars-6/Mars-7 descent module. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

After the end of the mission of the two Phobos spacecraft, an intense controversy erupted over whether another mission of the same type or a new mission should be developed whose primary goal would be to study the planet Mars itself. The latter point of view won out and led to approval of a very complicated project with the following five elements: a satellite in Mars orbit, small stations (landers), penetrators, a Mars rover, and an aerostat. The science experiments were largely developed within the framework of an international cooperative effort involving 20 countries. Launch was scheduled for 1994 but was then rescheduled for 1996. The situation was dramatic: Work on the project had begun in one country—the Soviet Union—and was continued and finally completed in another—the Russian Federation. The Russian Space Agency was established during this period, and Mars-96 became its first large scientific project. It soon became clear that it needed to be simplified. The Mars rover and aerostat were eliminated. However, there was a lack of funding, the project was continuously delayed, and the infrastructure crumbled. For the first time ever, the two-launch scheme had to be abandoned due to lack of funding.

Mars-96 was launched on 17 November 1996 but was unable to enter the transfer orbit because of a failure in the Block D stage. This was a catastrophe of a magnitude different from previous disasters. Many were convinced that this meant an end to the overall space-science research strategy inherited from the Soviet Union, with virtually unlimited funding, use of expensive launch vehicles, etc. However, it would have been very difficult for this strategy to continue, even

Table 1. Basic Specifications for Phobos-2 Spacecraft

Launch date	12 July 1988
Mars-orbit entry date	29 January 1989
Duration of Earth–Mars flight	200 days
Time spent in Mars space prior to Phobos flyby	120 days
Duration of Phobos flyby	25 minutes
Orbiter mass	2600 kg
Mass of scientific instrumentation	370 kg

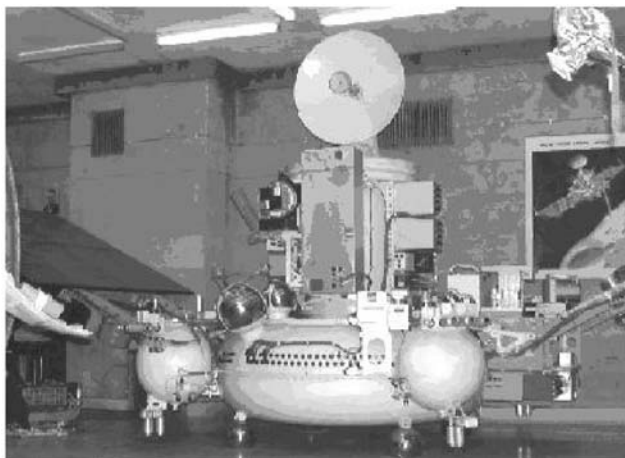
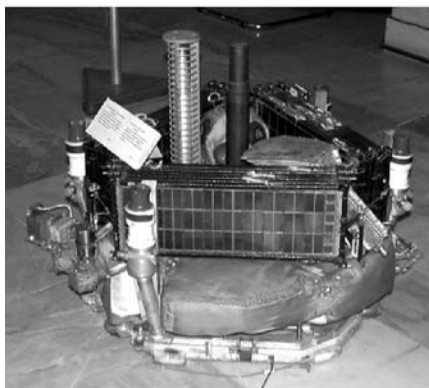


Figure 6. General view of the the Phobos-1/2 spacecraft. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

had the Mars-96 launch been successful. Basic data on the Mars-94/96 spacecraft and its scientific instrumentation may be found in two *Institute for Space Research* preprints (15,16).

While work on the Mars-96 project was under way, an attempt was made to organize bilateral cooperation between the United States and Russia in space-science research; this attempt was initiated by W. Huntress, who at that time was NASA Deputy Administrator for Space Science. The Cold War was now over, and it was logical to combine efforts in Mars research, as well as several other areas. Several options were generated for a joint flight to Mars under the title "Together to Mars": A Russian launch vehicle, an American orbiter, and a Russian descent module. For various reasons, this effort never proceeded beyond the preliminary discussion phase. This cooperative effort eventually merely amounted to participation of Russian scientists in three American Mars missions. Two of



Transfer configuration



Deployed configuration

Figure 7. General view of Phobos long-duration automatic station. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

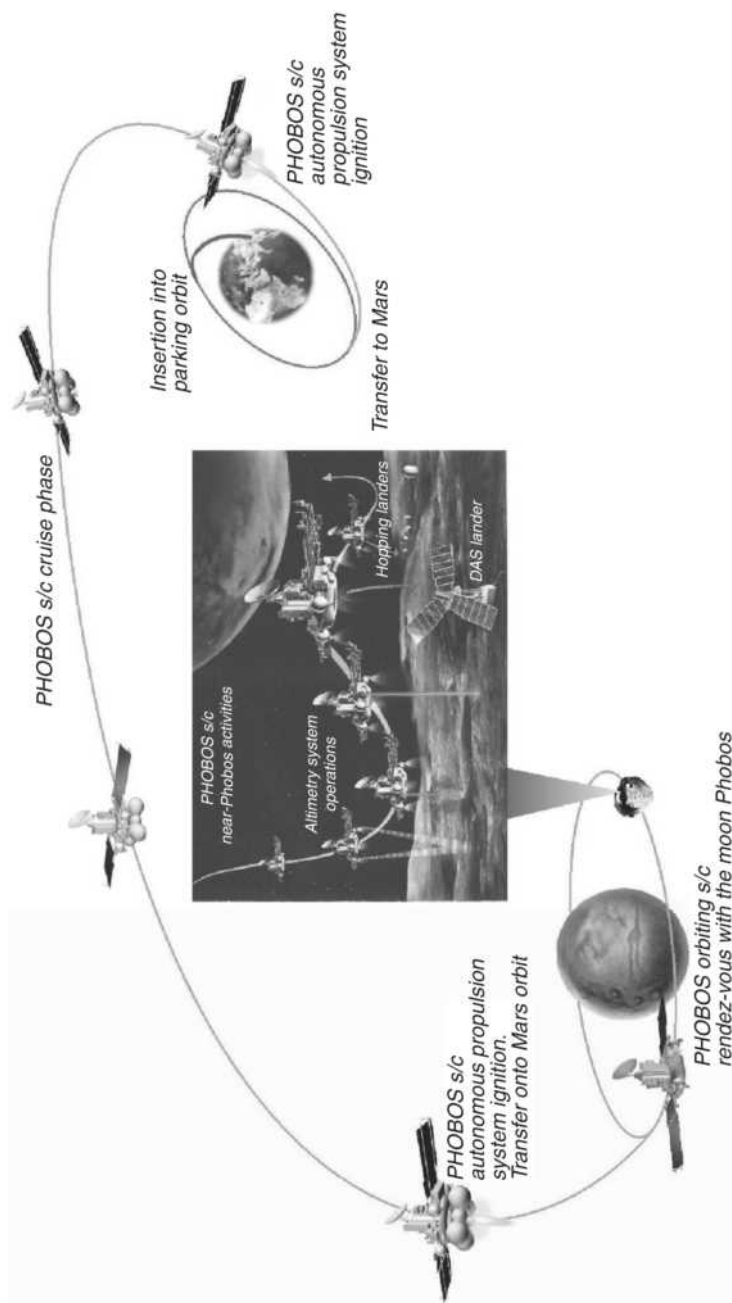


Figure 8. Phobos mission flight plan. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Table 2. Basic Specifications of Phobos Long-duration Automatic Station

Total mass	67 kg
Mass of scientific payload	20.6 kg
Height	1.8 m
Solar-panel length	2.5 m

them (Mars Climate Orbiter and Mars Polar Lander) failed but the third (Mars Odyssey 2001) was successful. Important observations of ice in the subsurface ground layer were made by Mars Odyssey with the gamma-spectrometer package that includes a Russian instrument (high energy neutron detector HEND).

Despite all of the lack of success in Mars research, Mars and its satellite, Phobos, remain highly attractive targets for the current Russian planetary program.

BIBLIOGRAPHY

1. Moroz, V.I. *Fizika Planet*. Nauka, Moscow, 1967 [*Physics of planets*, NASA TTF-515, 1968].
2. Chertok, B.E. *Rakety i Lyudi [Rockets and People]*, 2nd ed., Mashinostroenie, Moscow, 1999 [in Russian].
3. Chertok, B.E. *Rakety i Lyudi: Fili, Podlipki, Tyuratam [Rockets and People: Fili, Podlipki, and Tyuratam]*. Mashinostroenie, Moscow, 1999 [in Russian].
4. Chertok, B.E. *Rakety i Lyudi: Goryachie Dni Kholodnoi Voyny [Rockets and People: Hot Days of the Cold War]*. Mashinostroenie, Moscow, 1999 [in Russian].
5. Perminov, V.G. *The Difficult Road to Mars. A Brief History of Mars Exploration in the Soviet Union*. NASA, Washington, DC, 1999.
6. Series of papers on the results of the Mars-73 mission, *Kosmicheskie Issledovaniya* 13 (1): (1975) [in Russian].¹
7. Moroz, V.I. *Physics of the Planet Mars*. Nauka, Moscow, 1978 [in Russian].
8. Sagdeev, R.Z., V.M. Balebanov, and A.V. Zakharov. The Phobos Project. *Astrophys. Space Phys. Rev.* 6 (part 1): 3–62 (1988).
9. Zakharov, A.V. The plasma environment of Mars: Phobos mission results. In: J.G. Luhmann, M. Tatrallyay, and R.O. Pepin (eds), *Venus and Mars: Atmospheres, Ionospheres, and Solar Wind Interactions*. Geophysical Monograph Series #66, 1992, pp. 327–344.
10. Series of initial papers describing the results of the Phobos mission. *Nature*, 341: (1989).
11. Series of papers describing the results of the Phobos mission. *Geophys. Res. Lett.* 17 (6): (1990).
12. Series of papers describing the results of the Phobos mission. *Planetary and Space Sci.* 39 (1): (1991).
13. Sagdeev, R.Z., and A.V. Zakharov. Brief history of the Phobos mission. *Nature* 341: 581–585 (1989).
14. Snyder, C.W., and V.I. Moroz. Spacecraft exploration of Mars. In: H.H. Kieffer, B.M. Jakosky, C.W. Snyder, and M.S. Mattheus (eds), *Mars*. The University of Arizona Press, Tucson, 1992.

¹English translation of Kosmicheskie issledovaniya is available as Cosmic Research.

15. Zakharov, A.V. (ed.), Mars-94. *Inst. Space Res.* Preprint 1994.
16. Zakharov, A.V. (ed.), Mars-96. *Inst. Space Res.* Preprint 1996.

ALEXANDER V. ZAKHAROV
VASILY I. MOROZ
Space Research Institute
Russian Academy of Sciences
Moscow, Russia

EXPLORATION OF THE MOON BY SOVIET SPACECRAFT

Launches of Soviet spacecraft to the Moon started on 2 January 1959, only 15 months after the launch of the first Sputnik. The full list of missions to the Moon is given in Table 1. It includes all officially declared Soviet lunar missions. According to some sources there were also several failed Soviet attempts to send spacecraft to the Moon, when for various reasons, spacecraft stayed in Earth orbit or failed even earlier (1). They have never been officially recognized by the Soviet Union as lunar missions, and this is why they are not mentioned here. As a source for Table 1, we used the Soviet Encyclopedia of Cosmonautics (2). In total, there were 29 Soviet missions to the Moon. Of them, 20 were successful, one was partly successful, and eight missions failed. Twenty-four missions were named Luna (the Moon, in Russian); five missions were named Zond (probe, in Russian). All Soviet missions to the Moon were robotic, although there was a program of Soviet manned flights to the Moon, which was canceled in the late 1960s because of failures of the heavy booster rocket necessary for those flights (3).

It is necessary to mention that in the 1960s–1970s, when the Soviet Union sent spacecraft to the Moon, the space activity of both the Soviet Union and the United States was strongly politically motivated. The Soviet Union, by sending numerous lunar missions, wanted to demonstrate the superiority of its political system. The Apollo mission to the Moon, the backbone of the U.S. space program of that time, was a response to the challenge of Soviet space activity.

The first Soviet mission to the Moon, Luna 1, although it missed by 6000 km, was partly successful. It made geophysical measurements in near-Moon space and showed that the Moon does not have a significant magnetic field. Eight months later, Luna 2 reached the Moon and made scientific measurements until the spacecraft hit the surface. Data taken by magnetometer showed that the magnetic moment of the Moon could not be larger than 1/10,000 of the magnetic moment of Earth (4). The next Soviet lunar mission, Luna 3, October 1959, flew by the Moon sending back to Earth images of the far (invisible from the Earth) hemisphere of the Moon. Although the image quality was rather poor, it was discovered that the lunar farside, in comparison to the nearside, has much

Table 1. Soviet Missions to the Moon

Luna 1	Probe	2 Jan 1959
First attempt to reach the Moon. Missed it by 6000 km. Showed absence of lunar magnetic field. Sped up by lunar gravity and became the first artificial satellite of the Sun.		
Luna 2	Probe	12 Sept 1959
Successfully reached the Moon at 1° W, 30° N (435 km from the Moon's visible center), confirmed absence of lunar magnetic field.		
Luna 3	Probe	4 Oct 1959
Flew by the Moon and made first images of the farside of the Moon, invisible from Earth. The pictures were developed onboard and then sent back to Earth. Deficit of maria on lunar farside was discovered.		
Luna 4	Lander ?	2 Apr 1963
Mission failed: spacecraft missed the Moon by 8500 km.		
Luna 5	Lander	9 May 1965
Mission failed: spacecraft reached the Moon but due to a retrorocket failure crashed in Mare Nubium.		
Luna 6	Lander	8 Jun 1965
Mission failed: due to midcourse correction error, spacecraft missed the Moon by 160,000 km.		
Zond 3	Probe	18 Jul 1965
Flew by the Moon and made images of the farside and partly nearside of the Moon. The pictures were developed onboard and then sent back to Earth. Deficit of maria on lunar farside was confirmed.		
Luna 7	Lander	4 Oct 1965
Mission failed: spacecraft reached the Moon but due to a retrorocket failure crashed in Oceanus Procellarum.		
Luna 8	Lander	3 Dec 1965
Mission failed: spacecraft reached the Moon but due to a retrorocket failure crashed in Oceanus Procellarum.		
Luna 9	Lander	31 Jan 1966
First soft landing on the Moon (Oceanus Procellarum, 7.13°N, 64.37°W). First TV panoramas with close-up view of lunar surface were sent back to Earth.		
Luna 10	Orbiter	31 Mar 1966
First lunar satellite. Solar plasma sensors, magnetometer, micrometeorite sensors, gamma-ray spectrometer to measure surface composition of lunar surface.		
Luna 11	Orbiter	24 Aug 1966
Micrometeorite sensors, plasma sensors, gamma- and X-ray sensors to determine the Moon's chemical composition, tracking to measure lunar gravity field.		

Table 1. (Continued)

Luna 12 Onboard TV camera.	Orbiter	22 Oct 1966
Luna 13 Soft landing on the Moon (Oceanus Procellarum, 18.87°N, 62.05°W); TV panoramas sent to Earth; soil mechanics measured.	Lander	21 Dec 1966
Luna 14 Plasma and particle sensors; tracking to measure lunar gravity field.	Orbiter	7 Apr 1968
Zond 5 First lunar flyby with Earth return test; flight for Soviet manned expedition to the Moon, later canceled; a biological payload was included in the flight.	Flyby	14 Sept 1968
Zond 6 Lunar flyby with Earth return; continuation of test flight for Soviet manned expedition, micrometeorite and cosmic ray measurements, in-orbit photography, and biological payload.	Flyby	10 Nov 1968
Luna 15 Attempt of sample return from Mare Crisium; soft landing failed.	Lander	13 Jul 1969
Zond 7 Lunar flyby with Earth return; in-orbit photography of the Moon.	Flyby	7 Aug 1969
Luna 16 First successful robotic sample return; spacecraft landed in Mare Fecunditatis (0.68°S, 56.3°E).	Lander	12 Sep 1970
Zond 8 Lunar flyby with Earth return; in-orbit photography of the Moon.	Flyby	20 Oct 1970
Luna 17/Lunokhod 1 Soft landed in Mare Imbrium (38.17°N, 35°W); deployed the first robotic lunar rover which functioned for almost a year.	Lander	10 Nov 1970
Luna 18 Failed attempt of robotic sample return from lunar highland region in between Mare Fecunditatis and Mare Crisium.	Lander	2 Sep 1971
Luna 19 Micrometeorite, plasma and particle sensors, magnetometer, TV camera, tracking to measure lunar gravity field.	Orbiter	28 Sep 1971
Luna 20 Robotic sample return from lunar highland region in between Mare Fecunditatis and Mare Crisium (3.53°N, 56.55°E).	Lander	14 Feb 1972
Luna 21/Lunokhod 2 Soft landed in crater Le Monier at the eastern edge of Mare Serenitatis (25.85°N, 30.45°E); deployed robotic lunar rover which functioned for 4 months.	Lander	8 Jan 1973

Table 1. (Continued)

Luna 22	Orbiter	29 May 1974
Micrometeorite, plasma and particle sensors, magnetometer, TV camera, tracking to measure lunar gravity field.		
Luna 23	Lander	28 Oct 1974
Failed attempt of robotic sample return from Mare Crisium; the spacecraft was damaged on landing and could not function properly.		
Luna 24	Lander	9 Aug 1976
Robotic sample return from Mare Crisium (12.75°N, 62.2°E).		

less dark plains, called *maria* by astronomers. Now, this is, like the absence of a lunar magnetic field, common knowledge.

Then, the 3-year period of mission failures (Luna 4 to 8) in 1963 started, when the Soviet Union tried to soft land spacecraft on the Moon. This period was partly interrupted by the success of Zond 3, which flew by the Moon and sent back to Earth higher quality images of the lunar farside and part of the nearside, confirming the farside deficit in maria and providing knowledge of the photometric properties of the farside (5,6). Finally, in January 1966, Luna 9, successfully made a soft landing, 4 months before it was done by the U.S. Surveyor 1, and sent back to Earth TV panoramas of the close vicinity of the landing point (Fig. 1). This mission provided the first close-up view of the lunar surface (centimeter-size features were clearly seen). The successful landing had proved that the lunar surface is strong enough to withstand a load of this and future landers, a conclusion very important for exploration of the Moon (7).

Next, still in 1966, there was a series of three orbiters: Luna 10, 11, and 12. Luna 10 and 11 measured the micrometeorite and plasma environment in near-lunar space as well as gamma-ray radiation from the lunar surface. The latter is indicative of the chemical composition of the surface material. It was found that the surface material resembles basaltic lavas of Earth (8). At that time, these were the only measurements of the chemistry of the lunar surface. Luna 12 provided TV images of several areas of the lunar surface, some with a resolution as high as 15–20 m per pixel. Photogeologic analysis of these images led to a morphological classification of small (<1 km) lunar craters later used in the analysis of other images of the lunar surface. This series of successes was continued with the soft landing of Luna 13, December 1966, and the orbiter of Luna 14, April 1968. Luna 13, in its general design and presence of a panoramic TV camera was a repetition of Luna 9, but in addition, it had a mechanical penetrometer, dynamograph, and radiation densitometer to study lunar soil mechanics. It was found, in particular, that the surface material (at least 5 cm thick) consists of fine grains and its bulk density is $\sim 0.8 \text{ g/cm}^3$ (9). Luna 14 was making measurements of the micrometeorite and plasma environment, and its radio tracking (as well as tracking of Luna 10 to 12) was used to study the structure of the lunar gravity field that was required for better ballistic control of future lunar missions.

Then there was the flight of Zond 5, and a month later, Zond 6. Zond 5 was launched on 14 September 1968. Four days later it flew by the Moon and safely

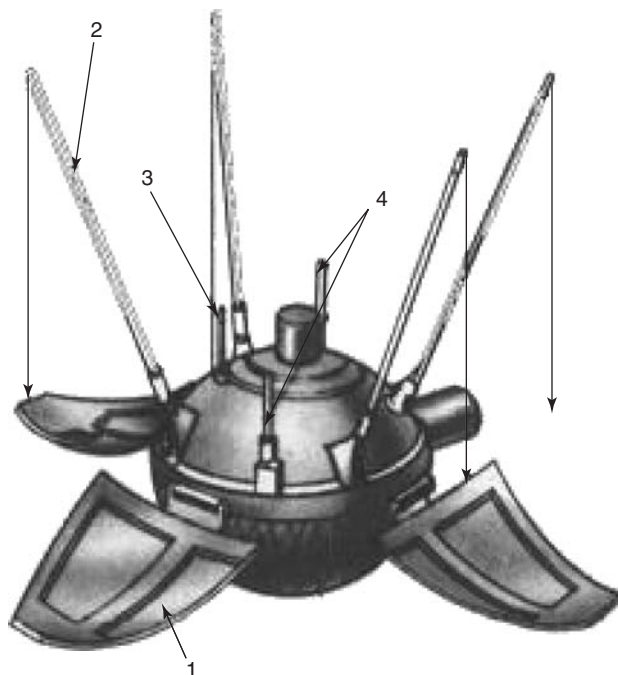


Figure 1. Luna 9 lander after opening its petals; 1-petal antennae; 2-rod antennae; 3-photometric standard; 4-two-edge mirrors. The lander mass 99 kg. A similar petal design, which guarantees that despite spacecraft orientation on landing it would take the correct (antennae up) orientation, was used in the U.S. Mars Pathfinder, which successfully landed 30 years later than Luna 9.

returned to Earth on 21 September. These were test flights of circumlunar modules of the Soviet manned mission to the Moon (at that time still in preparation). The science payload included micrometeorite and cosmic ray measurements, important for the safety of future crews, as well as in-orbit photography and a biological payload. The latter included turtles, wine flies, and other creatures that were the first living earthlings brought close to the Moon.

On 13 July, Luna 15 was launched, its mission was the first robotic return of lunar samples to Earth. It was a race with the Apollo 11 mission, which started on 16 July and landed American astronauts on the Moon on 20 July. Unfortunately, on 21 July, trying to land, Luna 15 crashed. The first successful robotic sample return was accomplished in September 1970, when Luna 16 landed in Mare Fecunditatis and brought back to Earth 101 g of lunar soil. In a year, after another failed attempt (Luna 18), Luna 20 successfully landed in the highland area between Mare Fecunditatis and Mare Crisium and brought back to Earth 55 g of lunar soil (Fig. 2). Then, again after a failed attempt (Luna 23), the last Soviet mission to the Moon, Luna 24, August 1976, successfully landed, drilled deep into the surface and brought the sample (a 1.7-m long drill core with 170 g total mass) back to Earth.

The returned lunar samples were thoroughly studied in Soviet and many other laboratories of the world and were partly exchanged with the samples brought by six Apollo expeditions. Thus, the Luna 16, 20, and 24 samples, despite

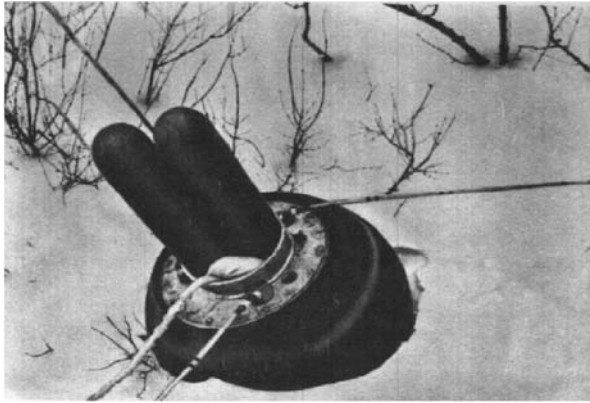


Figure 2. Returned capsule of Luna 20 with lunar samples inside landed in snows of Western Siberia. The two extended features in the top part of the capsule are inflated antennae sending signals to the recovery teams.

their relatively small mass, became an essential part of the research bank of lunar materials available for international scientific research that resulted in many publications (see e.g., (10–13)). Study of the Luna samples helped to construct a more complete picture of the distribution of different types of lunar materials along the visible part of the Moon. It was found, in particular, that the three Luna sites, all at the eastern equatorial part of the Moon, form a geologic province for which the presence of basalts, rich in aluminum and depleted in titanium and alkalis, was typical. It is necessary to add that studies of the Luna samples (and the Apollo samples as well) are still ongoing and produce more and more details in our understanding of our natural satellite.

The series of robotic sample returns was interrupted from time to time by other missions. The last two Zond missions were in August 1969 and October 1970 (7,8). They continued in-orbit photography of the Moon, covering, in particular, some areas of the farside not covered in necessary detail by the U.S. Lunar Orbiter and Apollo missions (14). There were also Luna 19 and 22 missions whose major goal was the study of the structure of the lunar gravity field. And finally, there were two more successful landings, Luna 17 and Luna 21, which brought to the lunar surface the research rovers, Lunokhod (in Russian: Moonwalker) 1 and 2.

Luna 17 brought Lunokhod 1 to the northwest part of Mare Imbrium. It had traveled 10,540 m, sent to Earth more than 50,000 pictures from the navigation TV cameras and more than 200 TV panoramas, conducted more than 500 lunar soil tests, and made numerous measurements of the chemical composition of the soil by the X-ray fluorescence technique. Lunokhod 1 also had a French-made laser retro reflector for high-precision measurements of the distances between the Moon and Earth. Luna 21 brought Lunokhod 2 to the mare-like surface of the floor of the large crater Le Monier at the eastern edge of Mare Serenitatis (Fig. 3). It had traveled 37,450 m, partly along the mare-like surface, partly intruding into hilly terrain of the highland type, and studying the edges of a 15-km long linear trough named Fossa Recta. It sent to Earth more than 80,000 pictures from the navigation TV cameras and 86 TV panoramas and conducted

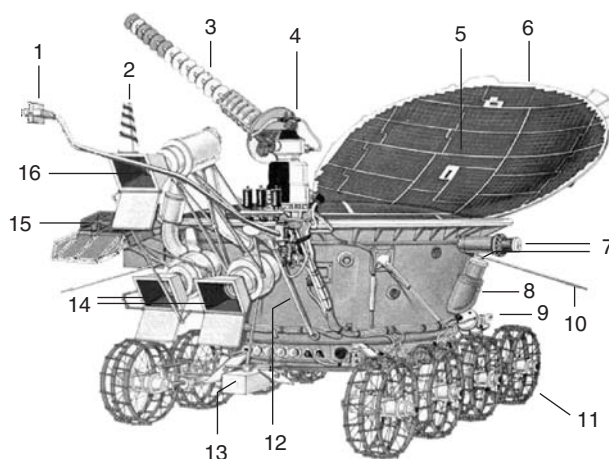


Figure 3. Lunokhod 2. General scheme; 1-magnetometer; 2-low-gain antenna; 3-high-gain antenna; 4-mechanism of antenna steering; 5-solar battery; 6-the cap (closed at nights); 7-panoramic TV cameras for vertical and horizontal scanning; 8-radioisotope heater; 9-instrument for soil-mechanics measurements; 10-rod antenna; 11-wheel with motor; 12-sealed instrument container; 13-X-ray fluorescence spectrometer; 14-stereo-scopic navigation cameras; 15-laser retroreflector; 16-upper navigation camera.

more than 150 lunar soil tests and numerous chemical analyses. In addition to that of Lunokhod 1, the Lunokhod 2 payload included a magnetometer, a photometric standard in the field of view of the panoramic TV cameras, and a special up-looking photometer to study the brightness of the night sky of the Moon.

The analysis of TV images taken by Lunokhod 1 and 2 led to a better understanding of the geologic processes responsible for the formation of the lunar soil and small-scale (centimeters to hundred meters) features of lunar topography (Fig. 4). In particular, it was found that surface regardening by meteorite



Figure 4. Fragment of Lunokhod 2 panorama showing rocky surface of the western edge of Fossa Recta trough and its opposite slope.

impacts was accompanied by a variety of down-slope gravity controlled mass-wasting phenomena (15–17). Joint consideration of local geology, the along the route measurements of the soil chemical composition (18) and soil optical reflectivity (19) led to conclusions on the mechanisms and scales of lateral and vertical mixing of lunar mare and highland materials. Measurements of the magnetic fields along the route and on the observation stations led to the discovery of local sites of residual magnetization associated with small impact craters and probably formed by the impacts (20,21). Analysis of variations of the interplanetary magnetic field registered by the Lunokhod 2 magnetometer led to estimates of the large-scale (hundreds of kilometers) structure of the Moon's interior (22). Lunokhod 1 and 2 measurements formed a very large database of lunar soil mechanics (23). Its analysis showed, in particular, that the soil is more cohesive on horizontal surfaces and less cohesive on steep slopes. Photometric observations of the lunar sky led to the discovery of a significant concentration of dust particles above the lunar surface (24).

As mentioned before, both Soviet and American lunar space programs were politically motivated. For about a decade, from January 1959 until December 1968, in the Soviet–American race to the Moon, there was approximate parity and several cases when “firsts” in key observations and discoveries were made by the Soviet Union. Since the Apollo program started its flights to the Moon, the U.S. role became dominant and some important achievements made by Soviet spacecraft lost importance because similar things had been done earlier and better by Apollo missions. The situation started to change after the last Apollo mission was completed (Apollo 17, 7–19 December, 1972). Meanwhile the Soviet studies of the Moon continued (Luna 21 to Luna 24), and it became obvious that the robotic approach to lunar studies could be very effective. The experience gained showed that Lunokhod type rovers could effectively study the Moon with a variety of instruments. Using the robotic arm, they could collect samples in places geologically interesting, but risky for landing, and reload them into the returning spacecraft of the Luna 16–24 type at sites safe for landing. The obvious advantage of this approach was that, compared to manned flights, it was much cheaper and could accept a much higher probability of risk. The Soviet lunar science community and space industry were actively discussing these possibilities, and even the Lunokhod 3 vehicle had been manufactured. But the Soviet leadership, which lost political interest in lunar studies, canceled all of these plans, and that was the end of Soviet missions to the Moon.

BIBLIOGRAPHY

1. Hart, D. *The Encyclopedia of Soviet Spacecraft*. Exeter Books, New York, 1987.
2. *Encyclopedia Kosmonavtiki (Encyclopedia of Cosmonautics)*, V.P. Glushko et al., (eds). Sovetskaya Encyclopedia Press, Moscow, 1985 (in Russian).
3. Chertok, B.E. *Rakety i Lyudi (Rockets and People)*. Mashinostroenie Press, Moscow, 2001 (in Russian).
4. Dolginov, Sh. Sh., et al. A study of the magnetic field of Moon. In *The Moon*, Z. Kopal and Z.K. Mikhailov (eds). IAU symposium 14, Academic, 1962, pp. 45–52.
5. Lipsky, Yu. N. Zond 3 photographs of the Moon's far side. *Sky and Telescope* 30: 338 (Dec 1965).

6. Lebedinsky, A.I., et al. Infrared spectroscopy of lunar surface from Zond-3 space probe. In *Moon and Planets II* A. Dollfus (ed.). North-Holland, Amsterdam, 1968, pp. 55–63.
7. *Pervye Panoramy Lunnoi Poverkhnosti*. Tom 1. Po materialam avtomaticheskoi stantsii Luna 9 (*First Panoramas of Lunar Surface*. Vol. 1. The results of robotic station Luna 9). Nauka Press, Moscow, 1967 (in Russian).
8. Vinogradov, A.P., et al. Gamma investigation of the Moon and composition of the lunar rocks. In *Moon and Planets II* A. Dollfus (ed.). North-Holland, Amsterdam, 1968, pp. 77–90.
9. *Pervye Panoramy Lunnoi Poverkhnosti*. Tom 2. Po materialam avtomaticheskikh stantsii Luna 9 and Luna 13 (*First Panoramas of Lunar Surface*. Vol. 2. The results of robotic stations Luna 9 and Luna 13). Nauka, Moscow, 1969 (in Russian).
10. *Lunnyi grunt iz Morya Izobilia (Lunar soil from Mare Fecunditatis)*. A.P. Vinogradov (ed.). Nauka Press, Moscow, 1974 (in Russian).
11. *Grunt iz materikobogo raiona Luny (Soil from the Highland Region of the Moon)*. V.L. Barsukov and Yu. A. Surkov (eds). Nauka, Moscow, 1979 (in Russian).
12. *Lunnyi grunt iz Morya Krizisov (Lunar soil from Mare Crisium)*. V.L. Barsukov (ed.). Nauka Press, Moscow, 1980 (in Russian).
13. Mare Crisium: The view from Luna 24. Proc. Conf. Luna 24, Houston, Texas, December 1–3, 1977. *Geochimica et Cosmochimica Acta*, Supplement 9. Pergamon Press, 1978.
14. Rodionov, B.N., et al. Relief of the moon's reverse side according to Zond 8 photographs. *Kosmicheskie Issledovaniya*, 14(4): 548–552 (1977).
15. *Peredvizhnaya Laboratoriya na Lune. Lunokhod-1 (Movable Laboratory on the Moon. Lunokhod-1)*. A.P. Vinogradov (ed.). Nauka, Moscow, 1971.
16. *Peredvizhnaya Laboratoriya na Lune. Lunokhod-1 (Movable Laboratory on the Moon. Lunokhod-1)*. A.P. Vinogradov (ed.). v.2. Nauka, Moscow, 1978.
17. Florensky, C.P., et al. A possible lunar outcrop—a study of Lunokhod 2 data. *The Moon*, 17: 19–28 (1977).
18. Kocharov, G.E., and S.V. Viktorov. Chemical composition of the moon's surface in the region of Lunokhod 2 operation. *Akademia Nauk SSSR, Doklady* 214: 71–74 (1974); *Soviet Physics-Doklady*, 19: 1–4 (1974).
19. Zasetskii, V.V., et al. Photometric studies of the lunar soil on the basis of Lunokhod-2 data. *Kosmicheskie Issledovaniya*, 19: 455–464 (May–June 1981).
20. Dolginov, Sh. Sh., et al. Study of magnetic field, rock magnetization and lunar electrical conductivity in the Bay Le Monier. *The Moon*, 15: 3–14 (Jan–Feb 1976).
21. Ivanov, B.A., et al. A possible magnetic field at the time of shock polarization of rocks as a possible cause of lunar magnetic-field anomalies associated with craters. *Pis'ma v Astronomicheskii Zhurnal* 2: 257–260 (May 1976) *Sov. Astron. Lett.* 2: 101–102 (May–June 1976).
22. Vanian, L.L., et al. Electrical conductivity anomaly beneath Mare Serenitatis detected by Lunokhod 2 and Apollo 16 magnetometers. *Moon and the Planets*, 21: 185–192 (Oct. 1979).
23. Leonovich, A.K., et al. The main peculiarities of the processes of the deformation and destruction of lunar soil. In NASA, Washington, *Soviet-American Conf. Cosmochemistry Moon and Planets*. Pt. 2, 1977, pp. 735–743.
24. Severnyi, A.B., et al. The measurements of sky brightness on Lunokhod 2. *The Moon*, 14: 123–128 (Sept. 1975).

ALEXANDER T. BASILEVSKY
Vernadsky Institute of Geochemistry
and Analytical Chemistry
Russian Academy of Sciences
Moscow, Russia

EXTRATERRESTRIAL LIFE, SEARCHING FOR

Introduction

Since at least the beginning of recorded history, humans have looked up at the sky and wondered whether or not we are alone. Traditionally, we have consulted philosophers and religious figures for guidance in answering this question. In every age and in every culture, these leaders have responded according to the particular belief system that they embraced at the time. Today, we live in a very special time. Beginning at the end of the twentieth century, humans acquired the knowledge and technological capability to try to answer this old and important question by doing experiments. As we begin the twenty-first century, scientists and the engineers may, in the near future, be able to present us with the first examples of life off planet Earth. At one extreme, we may be able to discern at a distance only some very suggestive biomarkers, without knowing whether they reveal the presence of microbes or minds. At the other extreme, it may be possible to discover another technologically advanced civilization and engage in some form of communication with it. In the middle ground, we may be able to acquire actual samples of biology from another world within our solar system and by returning them to our labs, unambiguously determine whether they represent an independent genesis of life. Or, to the contrary, whether all life on Earth can be traced to a distant place of origin.

Within this century, perhaps within our own lifetimes, we may know whether the laws of physics and chemistry, operating throughout the universe, routinely produce biology. Christian DeDuke has proclaimed that “life is a cosmic imperative” (1); we are poised to test that hypothesis.

Searching for Life Elsewhere in Our Solar System

Other Habitable Worlds. Liquid water is the key to all life as we know it. Water chauvinism will inevitably guide (and even bias) our search for habitable domains and life off Earth. This may represent an enormous failure of imagination, but it is the only practical approach to developing technologies and search strategies for life (and biology) as we do not yet know it. On the scale of previous human exploration, our own solar system is vast and continuously surprising; we cannot search everywhere at once. Choices need to be made and priorities established. For the next few decades at least, searches for extinct or extant life within our solar system will be a case of “follow the water.” During the exploration of those environmental niches where water exists or once existed, it will also be necessary to have a working definition of life so that we will know it if we find it. There is no perfect definition of life, but one that is used frequently is due to Gerry Joyce (2): “life is a self-sustained chemical system capable of undergoing Darwinian evolution.” Such a definition does not lend itself to exploration of our solar system by robotic surrogates. Instead, we will rely on evidence of liquid water, the presence of biogenic elements (C, H, O, N, S, P), and a

plausible source of free energy. If we can detect and document macroorganisms, the conclusions will be more straightforward. However, the expectation is that most of our searching will deal with microorganisms. At this small scale, even on Earth, the distinction between living and nonliving is not always clear.

More Than the Goldilocks Story. The concept of habitability, first introduced in Dole's *Habitable Planets for Man* (3), has been expanded into the habitable zone for life. This zone is the time-dependent volume surrounding the host star(s) over which liquid water could exist on the surface of a planet (4). Our own solar system makes it abundantly clear that in addition to stellar insolation, the atmospheric composition (the greenhouse gas components and the albedo) of the planets is another very important factor in enabling surface liquid water. Only recently have we begun to appreciate some of the other significant contributors to the Venus (too hot), Mars (too cold), and Earth (just right) Goldilocks story. The D/H ratio measured in the Venusian atmosphere and the dendritic channels photographed on the surface of Mars have convinced most scientists that 4 billion years ago, Venus and Mars boasted liquid water on their surfaces, as did Earth. As the primitive Sun evolved and brightened, Venus accumulated increasing concentrations of CO₂ in its atmosphere and underwent a runaway greenhouse that boiled away its water. Lightweight Mars lacked enough radioactive material in its interior to continue to keep it warm, and lost its surface water as runaway glaciation froze out its CO₂ atmosphere. On Earth, tectonic activity plays a crucial part in controlling terrestrial atmospheric CO₂ through the carbonate-silicate rock cycle (5). Breaking the cycle on Venus (dehydrated rocks became too brittle for subduction) and Mars (loss of energy source to drive plate motion) destroyed the stabilizing negative feedback loops. Subsurface liquid aquifers may still exist on Mars today and provide niches for the survival of life that originated and evolved during more clement times.

In 1996, the potential for life on early Mars was brought into sharp focus by claims that the ALH 84001 meteorite (a piece of rock ejected from Mars 16 million years ago and deposited on the Antarctic ice sheet 13,000 thousand years ago) contained fossil evidence for microbial life (6). Since that time, the debate has raged whether the material in question is of mineralogical or biological origin. The most recent assertion that tiny grains of aligned magnetite must have been produced by the analogs of magnetotactic bacteria (7) (because no abiological production mechanism is now known) has fueled the flames of this controversy, rather than settling the question. The rather recent realization that on Earth we have collected 16 samples of Mars (the so-called SNC meteorites) has reminded us that the terrestrial planets were not necessarily biologically isolated at the time when all three bodies were wetter and more conducive to the origination of life. If the answer for ALH 84001 turns out to be biology, the next question will be whether that biology is independent of, or related to, terrestrial biology. Might we be Martians?

Other Oceans under Ice. Images and data from spacecraft flying through or orbiting the Jovian system have caused us to expand the possibilities for the solar system habitable zone. Liquid water cannot survive on the surfaces of Jupiter's moons Europa, Callisto and Ganymede, but vast oceans may exist beneath their icy carapaces. The water would be kept liquid by the frictional heating produced from tidal distortions of the moons (alternately compressed and

stretched) as Jupiter and the other moons wage a gravitational tug of war throughout their orbits. This same energy source drives the extreme volcanism seen on the surface of Io. The evidence for these oceans is still indirect and consists of images of the fractured crusts that resemble familiar terrestrial ice fields floating over oceans (8), and magnetic field measurements that argue for a briny, conducting ocean beneath the ice of Europa (9). Spectral data provide evidence of organic structures on the surfaces of Callisto and Ganymede (and perhaps Europa) (10). Organics and liquid water have led scientists to begin speculating on free energy sources on Europa (and the other moons) capable of supporting metabolism for some form of life under the ice (11).

On Earth, we have been surprised by the deep, hot biosphere (12) beneath our feet, where microorganisms appear to exist in abundance, at depths up to 3 kilometers, independent of surface photosynthesis. Though life can apparently thrive at depth, we do not know whether life can evolve at depth. European oceans should prove to be a very interesting test case.

Missions of Exploration. The Taylor and McKee (2000) report on *Astronomy and Astrophysics in the New Millennium* (13) identified “Is there life elsewhere in the universe?” as one of five fundamental questions setting the scientific agenda for the decade to come. The committee further argued that within this millenium astronomers must “Search for life outside of Earth, and if it is found, determine its nature and distribution in the galaxy.” Thus in the United States, NASA has a clear mandate to undertake missions of exploration in search of life elsewhere. The Cornerstone missions of ESA reflect this same priority.

Although the Viking Landers found no trace of biology on the Martian surface, there has been growing interest in returning there to look for fossils or perhaps even extant life in any subsurface aquifers (14). Recently, the Mars Global Surveyor provided possible evidence for the existence of liquid water quite near the surface in the not so distant past (15). Previous mission failures have delayed, but not deterred, the continuing exploration of this close neighbor. Another orbiter, Mars Odyssey will be launched in 2001. Two surface rovers will be launched in 2003 along with ESA’s Mars Express. Placeholders exist for launches in the Mars Surveyor program in 2005 and 2007, and many planetary scientists are pushing for a sample return mission.

Europa will be the destination of an orbiter, originally scheduled for launch in 2003, whose goal is establishing whether a liquid water ocean does indeed exist beneath the surface ice. An affirmative answer will bolster the case for a lander to explore the ice in regions where water has recently reached the surface (and perhaps some form of hydrobot to probe the underlying ocean itself) for signs of life. It would be hard to overestimate the difficulty of this task. If life is distributed throughout the ocean and not strongly concentrated near the (presumed nutrient-rich) ice/water interface (11), the number of cells of microorganisms could be far less than 1 cm^{-3} .

The Cassini mission will drop its Huygens probe into Saturn’s moon Titan in 2004 to sample the organic chemistry there and perhaps provide valuable information about early Earth and chemical pathways to life. Bepi Colombo will be launched in 2009 to study the planet Mercury. While not directly relevant to biology, it will help us to understand better the conditions that lead to the formation of terrestrial planets.

The smaller bodies of the solar system will also get their share of attention. Radar observations from the newly upgraded Arecibo Observatory will explore the properties of nearby asteroids and their companions. NEAR Shoemaker has just ended its mission with a spectacular landing onto the surface of the asteroid Eros. This spacecraft has provided a wealth of data to help understand the conditions in the early solar system, but it has also posed surprising new questions. Very recently, the Pluto/Kuiper Express (planned to fly by Pluto in 2012 and go on to make a close approach to primitive Kuiper Belt objects) has been canceled. A more cost-effective substitute mission is now under study with the objective of reaching Pluto before 2020. Rosetta will be launched in 2003 to study the nucleus of comet Wirtanen. The Stardust Mission is on its way to collect interstellar dust and then return samples from the vicinity of Comet Temple II in 2006. As for a sample return from Mars, new analytical tools will be required to tease out any biological signatures from micro- and nano samples. As yet, there are no unambiguous methodologies for distinguishing terrestrial biology from an independent origin and evolution of life, but this challenge will be enthusiastically embraced should the sample returns demand it.

These missions are all subject to modification from fiscal pressures, equipment malfunction, and new scientific discoveries. The most reliable source of current mission planning can be found on NASA's Solar System Exploration Division web site at <http://solarsystem.nasa.gov/missions/mission.html> and on the ESA mission sites <http://sci.esa.int/home/missionsinprogress/index.cfm> and <http://sci.esa.int/home/futuremissions/index.cfm>.

Searching for Life Beyond Our Solar System

Other Habitable Planets. Life as we know it is a planetary phenomenon. The origin and evolution of terrestrial life were shaped by planetary circumstances, and life in turn has profoundly changed the planet on which we live. We are now extending our exploration for life within our solar system to the oceans of giant moons, and it is reasonable to reach farther and to search for life in the vicinity of other stars. To do so, we should begin by looking for habitable planets.

Searching for Planets. Planets shine in reflected starlight and are extraordinarily weak compared to their host stars. At optical wavelengths, the Sun outshines Earth by a factor of a billion to one. In the infrared at 10 microns, the contrast ratio is improved to about a million to one. Therefore, in principle, given an angular resolution that is sufficient to separate an image of the planet from an image of the star and methods for dealing with the extreme contrast ratio—adaptive optics, speckles, and nulling interferometers—any planets that may orbit the nearest stars could be directly imaged. That feat lies in our technological future. The methods of extrasolar planetary detection currently in use are all indirect.

It is possible to measure the reflex motion of a star about its planetary system's center of mass due to the gravitational pull of unseen orbiting planets. Astrometry measures this motion across the plane of the sky, and radial velocities measure the reflex along the line of sight to the star. Given a favorable alignment, one can also measure the diminution of stellar luminosity as a planet

transits in front of the star. Fortuitous alignment can also lead to magnification of the light of a distant star by a planet (sitting near the Einstein ring surrounding a foreground parent star) creating a short-lived gravitational microlensing event.

Measurements of reflex motions and direct imaging techniques will work best on nearby stars. The techniques also work best for massive planets and low-mass stars. Transits and microlensing events are transient and unpredictable and require long-term monitoring of large populations of stars. These techniques can provide planetary statistics and demographics for distant stars. The instrumental precision required for all these detection methods is extremely challenging (16), and in some cases require orbital platforms and new technological developments that will take one or more decades to achieve. In other cases, such as radial velocity studies, we have already achieved the required limiting precision, but we have not had the instruments on telescopes for the decade or more needed to detect planets in long-period orbits comparable to those of our own outer solar system. Recently Marcy et al. (42) reported the detection of a Jovian analog orbiting the star 55 Cancri.

Now, there are 99 reported extrasolar planets that orbit 86 stars and two pulsars (17). None of these is likely to be a habitable world. If the terrestrial mass planets that orbit the pulsar PSR 1257 + 12 ever were habitable, it is hard to conceive how they could have remained so during the supernova event that turned their host star into a pulsar. If the planets accreted after the event, they now exist in a hostile radiative environment without central warmth. As for the planets discovered orbiting normal stars, they are gas giants, often far more massive than Jupiter, and in surprisingly short-period orbits close to their host stars. These orbits are often very eccentric, and though the planets spend some time within the habitable zone of their stars, they also have extreme excursions from it. There is abundant scientific speculation whether these massive gas giants might have massive moons wrapped in thick atmospheres that might render them habitable. Testing this speculation about moons will be even more difficult than finding potentially habitable planets.

To date, we have found no analogs of our own solar system that have small terrestrial planets in short-period, circular orbits, and gas giants much farther out, in long-period orbits. We think that this is an observational bias. Observational tools currently lack the precision to detect terrestrial planets [except for transits that occur in eclipsing binary systems (18)], and we have not yet accumulated data long enough to detect the giant planets in wide orbits reliably. That will change within the decade. The radial velocity studies responsible for detecting the greatest number of planets have now achieved the 2 m/s precision required to yield an accurate measurement of planetary eccentricity (19). Therefore, it will be possible to detect any Jupiter analogs at orbital distances of ~ 5 AU and also to determine whether they actually travel in circular orbits. These will become the systems of choice for more intensive exploration for habitable planets.

What follows is a laundry list of projects and missions (now being implemented or planned for the future) that have potential for finding terrestrial-sized planets within the habitable zones of their stars. In general, no timelines are given because many of the spacecraft require very significant technological

development programs before they can be enabled. References to specific missions have been omitted, but they can be found (along with updated information on the projects) on two excellent websites in the United States and Europe: <http://exoplanets.org/> and <http://www.obspm.fr/encycl/encycl.html>.

Direct Imaging. Ground-based: nothing. Adaptive optics for the Kecks, Very Large Telescope (VLT), and other large aperture telescopes will permit imaging massive Jupiters, but cannot provide the contrast ratios needed for Earths.

Space-based: High resolution correction on the Next Generation Space Telescope (NGST) should allow detecting Earth-like planets around a nearest dozen stars within decade. The Planet Finder and Darwin will be free-flying nulling interferometers that work in the infrared and are being scaled to find Earth-like planets around the nearest 100 stars within two decades.

Astrometry. Ground-based: nothing. Narrowfield astrometry on the Kecks will yield 10 microarc second precision; good enough for Uranus, but not Earth (at a distance of 5 pc).

Space-based: GAIA should be able to detect 10-Earth-mass planets around the nearest 100 stars, and SIM promises a limit of 2-Earth-mass planets for the nearest 50 stars and 10-Earth-mass planets for another 200 stars. Both will provide finding lists for Terrestrial Planet Finder (TPF) and Darwin.

Radial Velocity. Ground-based: nothing. Both the HIRES and CORALIE spectrometers in time will be able to discover true solar system analogs by finding Jupiter-mass planets in large circular orbits. The other search techniques will then surely focus on trying to find any inner terrestrial planets in those systems.

Space-based: nothing. The fundamental limitation to the precision of this technique is set by the intrinsic fluctuations in the light from the star, so there is little to be gained by observing from above Earth's atmosphere.

Transits. Ground-based: The Transits of Extrasolar Planets (TEP) network of modest telescopes has achieved the precision to find 2-Earth-radii planets around a handful of nearby eclipsing binary stars. For these systems, it is assumed that the planetary orbits will be aligned with the stellar orbital plane, and therefore planetary transits are more probable. A network of dedicated telescopes is being proposed for detecting the gravitational micro lensing events caused by planets like Earth.

Space-based: Kepler will stare at 100,000 distant stars continuously for 4 years. It should achieve a photometric precision of 8×10^{-5} and detect Earth-sized planets in orbital periods of a year or less. It will also detect giant planets in short-period orbits. The mission expects to detect hundreds of terrestrial planets, but follow-up will be difficult because the population of stars is so far away. COROT will also spend one-half of its time in similar observations looking for planets in orbits of less than 0.5 AU.

Figure 1 illustrates the discovery space for extrasolar planets and the different observational techniques mentioned before, using some specific missions as examples.

Searching for Biomarkers. Compared to the other planets in our solar system, what most distinguishes Earth's atmosphere is its extraordinary disequilibrium chemical state. CH_4 coexists with O_2 . In equilibrium, these gases would

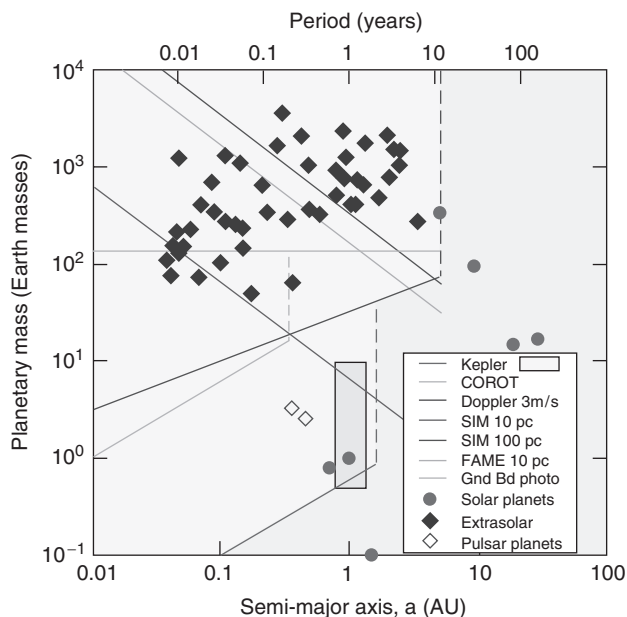


Figure 1. Observational Precision for Extrasolar Planet Searching Techniques (courtesy of Kepler website). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

quickly convert to CO_2 and H_2O . That conversion takes place in our atmosphere as well, but a continuous biological source function replenishes the trace gases; photosynthetic plants produce oxygen, and microbes in the guts of termites and bovine flatulence renew methane. In theory, an extremely sensitive, distant spectrometer working near 7 microns would be able to see the fingerprints of life as it viewed the absorption spectrum imposed upon the reflected sunlight by our atmosphere. This is a tremendous technical challenge. No spacecraft mission now on the books will have the required sensitivity to discover the trace levels of these two gases on any terrestrial planets that circle nearby stars. Our first attempts will be cruder and should also be less specific because life on other planets may more closely resemble early terrestrial methanogens before the invention of photosynthesis. Transits of short-period hot Jupiters offer the first opportunity to study potential biological markers in distant atmospheres. The CRIRES instrument on the VLT should be able to see CO , CH_4 , and H_2O in absorption during a transit, and NGST may be able to find the signature of chlorophyll absorption from 400–700 nm. Biology is not expected on gas giants themselves, but it will be most useful to know that organic materials are available in other planetary systems. When TPF and Darwin fly, they will attempt a spectral assay of any terrestrial planets they find. Predicted spectral sensitivity should be good enough to detect the broad absorption feature of ozone. If reasonable arguments can be made for the presence of liquid water on the planet's surface (so that the silicate–carbonate cycle would be functioning), then there is currently no known abiological process that could produce detectable amounts of ozone. If this all transpires several decades hence, it will be a circuitous way of

saying that evidence for some sort of biology had been found. What kind is another question.

Other Inhabited Planets. Perhaps the easiest way to detect other habitable planets will be by having the inhabitants themselves assist us. The field of SETI (search for extraterrestrial intelligence) does precisely that. SETI searches for various manifestations of another extraterrestrial technology. That technology might be used deliberately to announce the presence of an advanced civilization or for its own internal purposes. For more than a half century, Earth has been unintentionally announcing that we can broadcast radio and television signals. We are a very young technology in a very old galaxy (10 billion years old), so it is plausible to assume that there may be others. We do not know. We may never know. On the other hand, searches of increasing capability are now operating on about a dozen telescopes and could succeed between the time this article is written and published.

Searching for Technology. How can we possibly detect a distant technology? This question has been vigorously debated since the first paper was published by Cocconi and Morrison in 1959 (20), and Drake was preparing to conduct the first search in 1960 (21). A recent compendium (22) lists 99 searches published in the literature during the past four decades. The vast majority are searches for signals, and most of those have been at radio frequencies. It is possible to conceive of searches for physical artifacts, or miniaturized probes within our solar system, for enormous astroengineering projects detectable across interstellar distances, or indirect (remotely sensed) evidence that results from warfare on planetwide or planetary-system scales. It is harder to envision *systematic* search strategies for these types of evidence and harder still to devise experimental search protocols from which a negative result would be significant. These are the primary reasons that researchers have concentrated on looking for signals. Searches for the other types evidence are best left to serendipitous detection by an active observational program of astronomy at all wavelengths. However, this requires a willingness to consider technological explanations for newly detected phenomena that do not easily yield to astrophysical explanations.

Even searching for signals is a daunting task. The search space consists of at least nine dimensions: three space, one time, one frequency, two polarizations, one signal strength, and one modulation scheme. Figure 2 is a decision tree representing the various choices that must be taken to arrive at a particular search strategy. The shaded path represents the selections made by the Cyclops Report (23). Sponsored by NASA in 1971, Cyclops was the first engineering design study of ways to conduct a SETI program. Once the decision to search for electromagnetic signals has been made, it is necessary to select the frequency range; to decide whether to transmit and wait for a reply or to passively listen for signals already enroute; to decide whether to structure the search to look for some sort of leakage radiation or deliberate beacons; to decide whether to seek a signal continuously radiated toward Earth or one that will have some intrinsic duty cycle less than 100%; to decide whether to conduct the search with a single antenna or an array of elements; to decide whether to phase up the array or use the antennas independently; and to decide whether to use the antenna (or array) to study a single beam on the sky or to create multiple, simultaneous beams (in the extreme case, imaging the entire available primary field of view); to decide

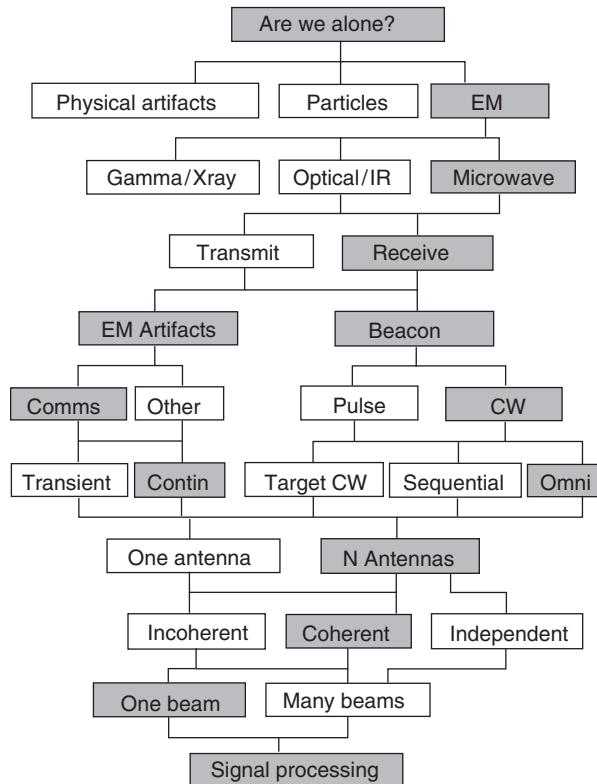


Figure 2. Decision tree for arriving at SETI signal search strategy.

whether to target individual stars or perform a sky survey; and finally, to decide what type of signal processing equipment needs to be acquired or invented to search for putative signals, given the other choices made at each level of the tree.

The signal processing step is the key to distinguishing technological from astrophysical signals. After many years of discussion, the basic criterion has remained unchanged. As far as we know, only technologies can produce a signal whose time-bandwidth product is close to unity ($B\tau \approx 1$), the limit set by the uncertainty principle. Because astrophysical emission processes are the result of very large collections of atoms, ions, or molecules, natural emitters will exceed this lower limit by many orders of magnitude (even in the case of intrinsically coherent processes such as masers). So search strategies and signal processing equipment have been developed to look for narrowband signals (continuous wave or long-duration pulses), and very short duration broadband pulses. Searches for CW and narrowband pulses at radio frequencies require extremely fine spectral resolution (resolving power $\sim 10^9$ – 10^{11}). Propagation through ionized interstellar medium disperses shorter radio pulses, eliminating the $B\tau \approx 1$ signature. At optical wavelengths, dispersion is not an issue, but absorption by interstellar dust becomes a limiting factor for distances beyond a kiloparsec. SETI searches for optical pulses require detectors that have time resolutions from nanoseconds to femtoseconds. Such devices have only recently become very affordable, and as a

result, searches at optical (though not yet infrared) wavelengths are beginning to offset the historical bias for radio wavelengths.

All searches must contend with both astrophysical and human-caused backgrounds. Much of the searching to date has been shaped by background considerations. To hear someone whispering in your ear, you would not choose to stand next to a waterfall but instead would listen in the quietest place in the forest. There are also waterfalls to be avoided in the sky. Figure 3 represents the natural astrophysical background radiation that any technologist within the Milky Way galaxy would encounter. Figure 3 assumes that the receiver is in space, without spectral, spatial, or temporal filtering. To date, almost every search has been done from the ground and each search has applied one or more types of filters. For a ground-based search, the atmospheric noise contributions and absorption from oxygen and water vapor need to be added to Fig. 3 at frequencies higher than 10 GHz.

Because we currently lack the spatial resolution to separate the emission from a transmitter on or near a planet orbiting another star from the stellar emission, any artificial signals must outshine the star to be detectable. At microwave frequencies this is an easy task. A star like our Sun is a weak radio source that produces only 1 kW/Hz. Thus, radio astronomy can be, and is, routinely conducted 24 hours per day, and at certain frequencies, our primitive technology handily outshines the Sun. The narrowband carrier waves embedded within broadcast FM and TV transmissions are factors of 10^6 to 10^9 brighter than the quiet Sun. Because stars like the Sun have a peak output at optical frequencies of $\sim 4 \times 10^{11}$ W/Hz, the power requirements are more severe for a continuous narrowband transmitter at those frequencies. Note that during a long

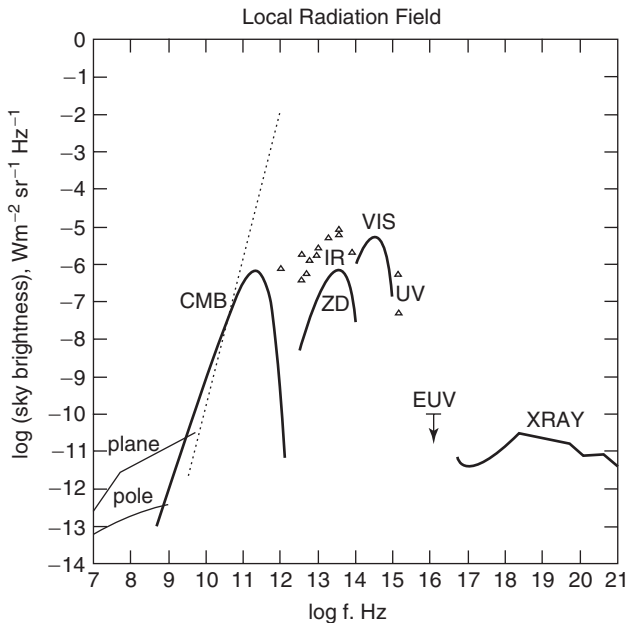


Figure 3. Background radiation from astrophysical sources.

observation, where integration is possible and the noise can be averaged down, the signal need only be detectable above the stellar fluctuations (rather than the total power). Rather (24) has argued that the transmitter power requirement is manageable for sufficiently narrow optical lines broadcast in the depths of the Fraunhofer absorption bands of the stellar spectrum.

SETI Today

Today, searches for fast optical pulses and CW signals as well as microwave searches for CW and narrowband pulses are all being conducted, using a variety of search strategies. Table 1. lists the currently active SETI searches that have committed to a regular program of observations. The search strategies, frequency coverage, sensitivity, and time on the air vary by many orders of magnitude within this table. Each of these projects maintains an active web site, and the reader is directed there for more details.

It is very difficult to compare the phase space searched by the various strategies represented in Table 1 and their relative effectiveness. Dreher and Cullers (25) attempted to define a figure of merit that will work for all SETI searches. They use the unknown transmitter power (continuous for CW signals and peak for pulses) as an explicit parameter. For each search strategy, they calculate the detection limit for an assumed transmitter power P_T and then they calculate the number of stars N_{stars} (foreground and background) that the search will explore. A logarithmic factor for the range of frequencies covered (from F_{lo} to F_{hi}) facilitates comparing optical and microwave searches on the same scale and yet favors the large frequency coverage of the former. A factor η_{pol} represents how thoroughly the signal polarization space is explored ($\eta_{\text{pol}} = 1$ for dual circular, 0.5 for single circular, and 1 for single linear with a corresponding reduction in the sensitivity by 0.5). The last factor is N_{looks} , the average number of times a star is observed in the course of the search program. A SETI figure of merit can then be given by

$$\text{FoM}(P_T) = N_{\text{stars}}(P_T) \times \ln(F_{\text{hi}}/F_{\text{lo}}) \times \eta_{\text{pol}} \times N_{\text{looks}}. \quad (1)$$

The exact form for the figure of merit is often debated; different emphases are placed on one or more of the terms. Figure 4 has been drawn assuming the preceding definition. It clearly indicates that targeted searches excel in finding faint nearby signals and sky surveys are superior in finding more powerful, distant transmitters. For sufficiently powerful transmitters, all sources become sky surveys seeing every star in the fraction of the galaxy that they cover. For more powerful sources still, these curves could be extended to include the neighboring galaxies and the Virgo supercluster. Whether such extraordinarily powerful transmitters exist is unknown. Figure 4 has been truncated within the Milky Way simply because there are insufficient data about the various search protocols to permit reliable predictions of their actual extragalactic coverage.

The 13 searches in Table 1 and Fig. 4 represent a higher level of global activity in SETI than at any time in the past four decades. This is surprising because all of the funding for these activities is now raised from private

Table 1. List of Active SETI Projects on Telescopes Today

DATE:	1990 – ON
OBSERVER(S):	LEMARCHAND “META II”
SITE:	Institute for Argentine Radioastronomy
INSTR. SIZE, m:	30 (one of two)
SEARCH FREQ., MHz:	1420.4, 1667, 3300
FREQ. RES., Hz:	0.05
OBJECTS:	Sky survey of southern skies and 90 target stars, and OH masers
FLUX LIMITS, W/m²:	1×10^{-23} to 7×10^{-25}
TOTAL HOURS:	Continuing
REFERENCE:	http://www.planetary.org/html/UPDATES/seti/META2/default.html
COMMENTS:	Search for signals that have been Doppler compensated to rest frame of SS barycenter, Galactic Center, or CMB. A duplicate of META system build by Argentinian engineers under the guidance of Prof. Horowitz at Harvard and financed by the Planetary Society. Simultaneous observations with META over declination range -10° to -30° . Major upgrades in 1996 to permit long integration times and switching between antennas. Search through OH masers looking for amplified signals.
DATE:	1995 – ON
OBSERVER(S):	KINGSLEY
SITE:	COLUMBUS OPTICAL SETI OBSERVATORY, OHIO
INSTR. SIZE, m:	0.1
SEARCH FREQ., MHz:	0.55 microns
FREQ. RES., Hz:	none
OBJECTS:	Nearby solar-type stars
FLUX LIMITS, W/m²:	Transmitters with peak instantaneous power $> 10^{18}$ W
TOTAL HOURS:	Ongoing
REFERENCE:	http://www.coseti.org
COMMENTS:	Broadband optical search for short pulses (~ 1 ns) that instantaneously outshine the host star. No formal program of observation is currently being conducted due to equipment damage.

DATE:

OBSERVERS:

SITE:

INSTR. SIZE, m:

SEARCH FREQ., MHz:

FREQ. RES., Hz:

OBJECTS:

FLUX LIMITS, W/m²:

TOTAL HOURS:

REFERENCE:

COMMENTS:

1995 – ON

Horowitz ET AL. (BETA)

Oak Ridge Observatory

26

1400 to 1720

0.5

Sky Survey from -30° to $+60^\circ$ declination
 2.2×10^{-22}

Suspended in spring 1999

<http://mc.harvard.edu/seti/beta.html>

Waterhole search, using dual beams and omni antenna to discriminate against RFI. Project BETA is following on to META. Project interrupted when wind blew antenna off its mount. Repairs are underway.

DATE:

OBSERVERS:

SITE:

INSTR. SIZE, m:

SEARCH FREQ., MHz:

FREQ. RES., Hz:

OBJECTS:

FLUX LIMITS, W/m²:

TOTAL HOURS:

REFERENCE:

COMMENTS:

1996 – ON

SETI LEAGUE PROJECT ARGUS

MULTIPLE SITES WORLDWIDE (currently ~ 100)

~ 3 – 10 (satellite TV dishes)

1420–1720

1

All sky

$\sim 1 \times 10^{-21}$ (varies)

Ongoing

<http://www.setileague.org>

Plan to organize up to 5000 radio amateurs to provide continuous sky coverage for strong, transient signals using systems that can be bought and built by individuals. SETI League currently has 1345 members running 115 sites.

Table 1. (Continued)

DATE:		1996 – ON
OBSERVERS:		WERTHIMER ET AL. (SERENDIP IV)
SITE:		Arecibo
INSTR. SIZE, m:		305
SEARCH FREQ., MHz:		1420 ± 50
FREQ. RES., Hz:		0.6
OBJECTS:		Survey of 30% of sky visible from Arecibo
FLUX LIMITS, W/m²:		5×10^{-24}
TOTAL HOURS:		Ongoing
REFERENCE:		http://seti.ssl.berkeley.edu/serendip/serendip.html >
COMMENTS:		Commensal search occurring at twice sidereal rate in backward direction while radio astronomers track targets using Gregorian system. Covers sky every 3 years, looks for signals recurring at same frequency and location on rescan.
DATE:		1998 – ON
OBSERVER(S):		SETI Institute Project Phoenix
SITE:		Arecibo Observatory and Lovell Telescope at Jodrell Bank
INSTR. SIZE, m:		305 m and 76 m
SEARCH FREQ., MHz:		1200 to 3000 dual pol
FREQ. RES., Hz:		1
OBJECTS:		600 nearby stars
FLUX LIMITS, W/m²:		1×10^{-26}
TOTAL HOURS:		1300 hours to date
REFERENCE:		http://www.seti.org >
COMMENTS:		Continuation of NASA HRMS targeted search of 1000 nearby stars, using real-time data reduction and a pair of widely separated observatories to help discriminate against RFI.

DATE: 1998 – ON
OBSERVER(S): SETI Australia Southern SERENDIP
SITE: Parkes
INSTR. SIZE, m: 64 m
SEARCH FREQ., MHz: 1420.405 ± 8.82
FREQ. RES., Hz: 0.6
OBJECTS: Southern sky survey
FLUX LIMITS, W/m²: 4×10^{-24}
TOTAL HOURS: Ongoing
REFERENCE: <http://seti.uws.edu.au> >
COMMENTS: Comensal search that uses 2 out of 13 beams of Parkes focal plane array to discriminate against.

DATE: 1998 – ON
OBSERVER(S): Werthimer (SEVENDIP)
SITE: Leuschner Observatory
INSTR. SIZE, m: 0.8
SEARCH FREQ., MHz: 300–650 nm
FREQ. RES., Hz: None
OBJECTS: 800 solar-type stars
FLUX LIMITS W/m²: 1.5×10^{-9} peak during 1 ns pulse, or 1.5×10^{-20} average per 100 second observation
TOTAL HOURS: 200 (ongoing)
REFERENCE: <http://sag-www.ssl.berkeley.edu/opticalseti> >
COMMENTS: First optical search to use two high time resolution photodetectors in coincidence to look for nanosecond pulses.

Table 1. (Continued)

1998 – ON	
OBSERVER(S):	Horowitz et. al. Harvard Optical SETI
SITE:	Oak Ridge Observatory
INSTR. SIZE, m:	1.5 m
SEARCH FREQ., MHz:	350–700 nm
FREQ. RES., Hz:	None
OBJECTS:	13000 solar-type stars of which 4000 done to date
FLUX LIMITS, W/m²:	4×10^{-9} peak in <5-ns pulse, or 4×10^{-20} average per 500-second observation
TOTAL HOURS:	Ongoing
REFERENCE:	http://mc.harvard.edu/oseti/index.html
COMMENTS:	Search for nanosecond laser pulses, with hybrid avalanche photodiodes in coincidence. Piggybacks on nightly searches for extrasolar planets.
1998 – 2000	
OBSERVER(S):	Marcy, Reines, Butler, Vogt
SITE:	Lick, Keck
INSTR. SIZE, m:	10 m
SEARCH FREQ., MHz:	400–500 nm
FREQ. RES., Hz:	Resolving power = 50,000
OBJECTS:	600 FGK stars within 100 pc
FLUX LIMITS, W/m²:	1×10^{-13}
TOTAL HOURS:	500
REFERENCE:	http://albert.ssl.berkeley.edu/opticalseti
COMMENTS:	Search through archival data for narrowband continuous optical laser emission lines

DATE:	1999 – ON
OBSERVER(S):	WERTHIMER AND ANDERSON (SETI@HOME)
SITE:	ARECIBO
INSTR. SIZE, m:	305
SEARCH FREQ., MHz:	1420.405 \pm 1.25 MHz
FREQ. RES., Hz:	0.6 Hz
OBJECTS:	Data taken from SERENDIP IV—sky visible from Arecibo
FLUX LIMITS, W/m²:	5 \times 10 ^{−25}
TOTAL HOURS:	Ongoing
REFERENCE:	〈http://setiathome.ssl.berkeley.edu〉
COMMENTS:	Hugely successful experiment in distributed computing. Permits more sophisticated processing of a fraction of SERENDIP IV data by harnessing idle CPU cycles of 2.8 million personal and corporate computers.
DATE:	2000 – ON
OBSERVER(S):	MONTEBUGNOLI (SETI Italia)
SITE:	Medicina
INSTR. SIZE, m:	32
SEARCH FREQ., MHz:	1415–1425 and 4255–4265
FREQ. RES., Hz:	0.6
OBJECTS:	Northern sky
FLUX LIMITS, W/m²:	No routine observing program established yet
TOTAL HOURS:	Ongoing
REFERENCE:	〈http://www-radiotelescopio.bo.cnr.it/setiweb/home.htm〉
COMMENTS:	Commensal sky survey using Medicina telescope and SERENDIP signal processing boards.
DATE:	2000 – and On
OBSERVER(S):	Bhathal and Darcy
SITE:	Campbelltown Rotary Observatory , Oz OSETI
INSTR. SIZE, m:	0.4 and 0.3
SEARCH FREQ., MHz:	550 nm
FREQ. RES., Hz:	None
OBJECTS:	200 solar-type stars
FLUX LIMITS, W/m²:	200 southern stars and Additional globular clusters
TOTAL HOURS:	Ongoing
REFERENCE:	〈http://www.coseti.org/ragbir00.htm〉
COMMENTS:	Dedicated telescopes built for SETI. Uses high time resoution photodiodes in coincidence to search for laser pulses, soon to use two telescopes in coincidence and to be teamed with microwave search of same objects.

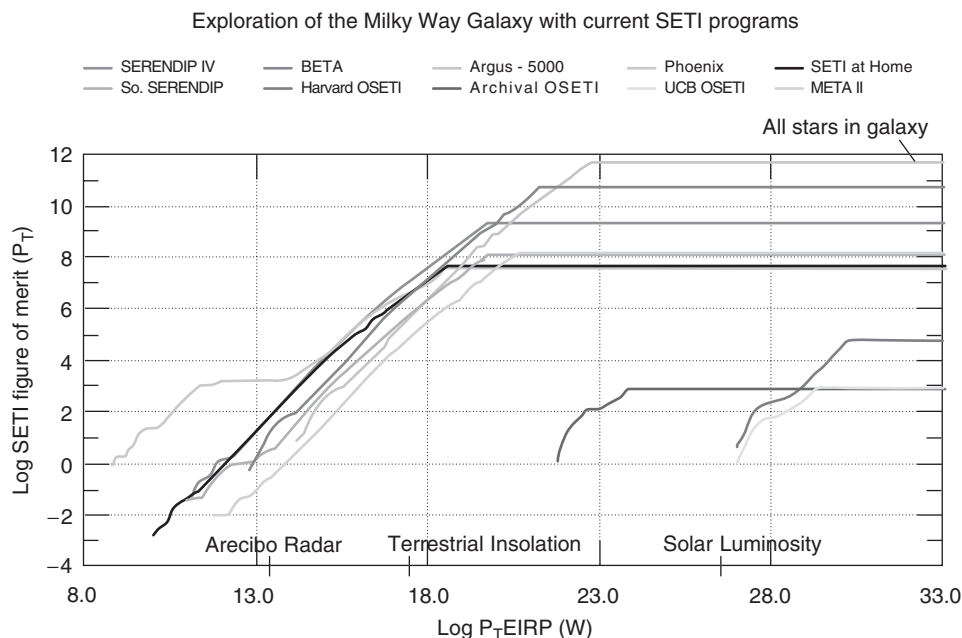


Figure 4. Exploration of the Milky Way Galaxy with current SETI programs. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

(nongovernmental) sources. From the mid-1970s to 1 October, 1993, NASA managed a SETI program named (finally) the High Resolution Microwave Survey (HRMS). This project consisted of two complementary search strategies: a targeted search from 1–3 GHz that examined 1000 nearby solar-type stars using large radio telescopes that belonged to the global scientific community and an all sky survey from 1–10 GHz enabled by the multiple 34-m antennas of the NASA Deep Space Network. Although there was some SETI activity in other countries, the NASA project and the United States were the primary players. Since Congressional termination of the HRMS, the level of private funds raised from philanthropic sources has (of necessity) grown dramatically. The largest project is the SETI Institute's Project Phoenix that uses modified equipment derived from NASA's HRMS to continue the targeted search. Jointly, the Planetary Society and the SETI Institute funded expansion of the SERENDIP project to survey one-third of the sky using the Arecibo Observatory across a reduced frequency range (compared to HRMS). Australia and Italy have now created privately funded support organizations to move SETI forward in those countries and have active search projects running commensally on large radio astronomy facilities and dedicated optical telescopes. By far, the most impressive indication of global interest in SETI is the success of the SETI@home experiment in distributed computing. More than 3.8 million people in more than 226 countries have downloaded this screen saver to assist in SETI data processing. The SETI League Project Argus hopes to establish a global organization of 5000 amateurs using small antennas to cover the sky at all times, looking for powerful transient sources. To date, 105 stations are operating in 19 countries.

SETI Tomorrow

Against this backdrop of global interest and reliance on philanthropic funding, the SETI Institute convened a series of workshops (from 1997–1999) to develop a road map for the next two decades, trying to predict what would be technologically feasible and affordable. The SETI 2020 Report (26) from those workshops recommends three projects:

1. An large array of small antennas that have 10^4 square meters of collecting area, that cover a factor of 20 in microwave frequencies, and can conduct SETI and traditional radio astronomy simultaneously by multibeaming and imaging. They called this the One Hectare Telescope or 1hT.
2. Initiating optical SETI searches for nanosecond pulses, using two high-time-resolution photodiodes in coincidence to reduce the false alarm rate from photons generated in the optics and detectors. It was envisioned that the first searches would take place on modest telescopes in the university community, to move down in frequency into the IR when detectors became available (to avoid extinction due to interstellar dust), and to migrate onto larger collecting areas as the field matured.
3. An omnidirectional SETI system (OSS) for the lower portion of the microwave window, following the early ideas for a radio camera first suggested for SETI by Dixon (27). The OSS is intended to enable searches for strong but transient signals (such as might be expected from a phased array beacon sequentially illuminating a large list of targets) or emission from sporadic astrophysical sources. The computational requirements for this equivalent of a radio-fly's eye are extreme because it must form all possible beams on the sky from horizon to horizon and provide high spectral resolution to search for signals. For an array of 4096 dipole elements, an estimated 10^{16} operations per second would be required. Although not available today at any price, this real-time computing capacity should be affordable in approximately 15 years. The workshop recommended a plan starting with an array of 4×4 elements and 4 MHz of bandwidth, growing exponentially as Moore's law improvements permit.

As the entries in Table 1 attest, optical SETI (OSETI) was enthusiastically embraced. Programs were operating on telescopes at UC Berkeley and Harvard, using target lists of several thousand stars, before the workshops concluded. Since then, OSETI projects have been started at Princeton (to work in coincidence with Harvard), at Lick Observatory, and at a newly constructed facility in Australia. Ground has been broken for an innovative optical sky survey telescope at Harvard that can survey the 80% of the sky visible from that site in about 150 clear nights. This survey will eventually cover the million stars within 1000 light-years and the space between, looking for strong signals. Initial discussions have begun to find ways that OSETI experiments can be piggybacked onto new Cherenkov detection arrays that consist of large light buckets and focal plane arrays of photodiodes (or other suitable experiments) to increase OSETI sensitivity in the most cost-effective way. OSETI detectors on 1-m-class telescopes

today could detect a nanosecond petawatt-pulse (1 megajoule) focused by a 10-m telescope from a distance of 200 light-years.

The concept of the 1hT has also been adopted. In partnership with the Radio Astronomy Laboratory at UC Berkeley, the SETI Institute is conducting the technology development for an array to be built in northern California at the Hat Creek Observatory. The Paul G. Allen Foundation has provided the development funding for what is now called the Allen Telescope Array (ATA), and Nathan Myhrvold has funded the electronics laboratory for the facility. The array consists of 350 offset-Gregorian antennas, each 6 m in diameter, with a 2.4-m subreflector, that are manufactured by the same hydroforming process used to mass-produce TVRO antennas. Construction should begin in 2003, following two preliminary prototyping stages necessary to validate the performance of the emerging consumer technologies now being adapted to provide unprecedented wideband performance and low cost. The array should commence operation in 2005 using three independent, synthesized beams on the sky, 100 MHz of processing bandwidth at frequencies tunable from 0.5–11 GHz, and an imaging processor of comparable bandwidth for the large field of view of the ATA (28). If funding can be found, the ATA will begin to expand antennas, beams, and bandwidth immediately. In so doing, the ATA will increase its sensitivity and serve as a proof-of-concept for the construction of an array that has 100 times the ATA initial collecting area—the Square Kilometer Array (SKA). This international project has as its goal 100 steerable beams on the sky, a frequency range from 150 MHz to 20 GHz, nanoJansky-level sensitivity, and a million-to-one imaging dynamic range. At present, five different concepts are being studied for the SKA in 10 different countries (29). To achieve the ambitious straw-man science requirements, a hybrid approach based on two or more of the current concepts will probably be required. However it is constructed, the SKA will improve the current sensitivity of microwave SETI searches [that can now detect strong terrestrial radars (10^{12} watts EIRP) at 150 light-years] by two orders of magnitude and open the door for searches of weaker “leakage” radiation from the vicinity of the nearest stars.

SETI 2020 made it clear that it would be more than a decade before the OSS becomes affordable. Nevertheless, a small-scale prototype array for the OSS is currently being developed at Ohio State University in collaboration with the SETI Institute. The Argus telescope, as it is called, will permit development of beam-forming, signal processing, and radio frequency interference excision algorithms, even though its sensitivity will remain low until the number of array elements (and thus the real-time computational loads) can be increased significantly.

All of these new searches have to contend with artificially high false alarm rates. At optical wavelengths, high-energy particles and coronal discharges in the detectors themselves are the problem, requiring coincidence detection schemes. Even using two photodiodes in coincidence, the spurious detection rate has remained unacceptably high (~ 1 per night); additional efforts are required. At Lick Observatory, three photodiodes have recently been used in coincidence to beat down the false alarm rate to nearly zero. The groups at Harvard and Princeton have collaborated to operate those two telescopes (each with two photodiodes) in coincidence, using GPS to synchronize the systems to 20 microseconds across the 1.6-millisecond path difference between the sites (30). To date, there are no

human-produced flashes of nanosecond light on the sky, but this could change in the future and would place high-time-resolution optical astronomy and SETI in the same situation that now disrupts microwave searches.

At microwave frequencies, the spurious detections result from our own use of the spectrum for all manner of communications. Even though microwave transmitters may be operating within their allocated frequency bands, from the view point of a comprehensive microwave SETI search, such signals represent undesirable radio-frequency interference (RFI). This is a worsening problem for all of the passive services as the number of satellite constellations increases. In addition, profit incentives often result in inadequate filtering and thus enhanced out-of-band and spurious emissions throughout the spectrum. Geographical isolation is no longer adequate to avoid RFI. The Phoenix project routinely conducts its targeted search in real time using two widely separated antennas (currently the 1000-foot Arecibo Observatory in Puerto Rico and the 250-foot Lovell telescope in the U.K.). Demanding detection of a narrowband signal at both sites with the correct differential Doppler signatures (31) significantly decreases the occurrence of false positives by discriminating against signals local to one site, as well as overflying or orbital platforms whose footprint encompasses both sites. The ATA will elaborate on this basic scheme by using the natural decorrelation of the signals across the extended array, by actively steering null beams of the array pattern to follow individual satellites, and by using one or more of the individual dishes within the array to observe the interfering transmitter with sufficient interference-to-noise ratio to permit accurate subtraction from the output of the array.

Before the SKA can be successfully implemented, serious consideration must be given to establishing an international radio quiet zone (IRQZ) that encompasses at least its central concentration of elements. No such zones exist today, nor is there any precedent for establishing such a zone internationally. Nevertheless, if the SKA can be sited in an area of sufficiently low population density (where the potential market is small), it may be possible to negotiate agreements with service providers so that they modify or turn off their transmitter beams when overhead. If this fails, some frequencies will be permanently unavailable to SETI and radio astronomy for some, or all, of the time. The only alternative would then be to construct observatories within the "Shielded Zone of the Moon" defined by the international ITU treaty in 1979 (32). Compared to the costs of doing business on the lunar farside, extensive investment in active strategies for RFI mitigation, alternate service delivery systems for affected populations, and even negotiations for provider compensation are likely to be cheaper.

What If the Searches Succeed?

The detection of microbial life beyond Earth would be spectacular. If it represents an independent origin and evolution, then it will revolutionize biology in all of its forms. We will begin to distinguish the necessary properties of life from the merely contingent. If we discover intelligent life elsewhere, it may be possible to shortcut the period of our technological infancy and understand our Universe and our place therein more rapidly than we would have done in isolation. In either case, positive results carry with them significant challenges.

Sample Return and Planetary Protection. On 15 January 2006, a capsule from the Stardust mission will parachute down over Utah returning samples of dust from the coma of Comet Wild 2. This is the first time since the Apollo Moon landings that extraterrestrial material samples will have been returned to Earth by spacecraft. However, because a billion dust particles from comets and asteroids fall to Earth every second, it is hardly the first time cometary dust has arrived here. The sample will contain about 1000 intact particles that were slowed, captured, and spike-heated in aerogel, and millions more fragments totaling only a few micrograms in mass. Although science fiction has considered the possibility of life within a liquid water reservoir deep inside a comet, the Planetary Protection Task Group of the National Research Council has concluded that there is no possibility that life forms will be returned in the Stardust cometary dust samples (33) and no hazard protection is required. The same cannot be said for a sample return from Mars, likely to occur in the second decade of the twenty-first century. In that case, the Task Group on Issues in Sample Return (34) recommended treating the returned samples as potentially hazardous, until proven otherwise, and curating them in a special sample-receiving facility. Given the desire to avoid both false positives and any alteration of the Martian environment, the Task Group also considered the problems of forward contamination from Earth to Mars and the levels of sterilization required for any spacecraft. These latter concerns will also be relevant to any future mission to sample probable liquid oceans beneath the ice crusts of Jupiter's large moons Europa, Callisto, and Ganymede. NASA's Planetary Protection Officer coordinates with the Planetary Protection Panel of COSPAR to internationalize standards and mission plans. Certainly, the global resources of the scientific community will be required to tease out the subtle distinctions between life versus nonlife and life-as-we-know-it versus exobiological life from the micrograms or nanograms of any future sample returns.

Our First Trip to the Stars. Interstellar travel is not possible now, and it is likely to remain uneconomic except for miniaturized robotic probes for the foreseeable future. Advanced propulsion systems are being actively studied by NASA at JPL (35) and MSFC (36) and at comparable facilities within ESA. For near-term interstellar precursor missions, solar sails and high-powered electric propulsion are being developed. Direct detection of terrestrial planets around stars within tens of light-years from Earth by TPF or Darwin and particularly, the remote sensing of biomarkers in the planetary atmospheres will enhance our resolve to reach the stars. We will want to send our surrogates there to view the beachfront property and look for other signs of life, as the Galileo spacecraft observed Earth during its 1990 and 1992 flybys (37). Whether this will be enough of an incentive to overcome the financial hurdles inherent in developing the new propulsions systems and information technologies necessary for such missions remains to be seen.

Who Will Speak for Earth? What Will They Say?

Any technological civilization that we detect will be more advanced than we are—they could not be less advanced—or they would not possess any technology that could be recognized over interstellar distances. Because we are a

hundred-year-old technology in a ten-billion-year-old galaxy, the odds that they are at a comparable technological level are tiny. The Drake equation (38) (with all of its uncertainty) tells us that technological civilizations will not be close to us in both space and age. If they are nearby, they will be old, if they are very distant, they could be as young as we are. In the former case (they are nearby and old), our first encounter with them is unlikely to be their first contact. They will have done this all before. We can expect a deliberate signal whose message content is intended to be anti cryptographic; in time we might even understand it. This scenario (so popular in science fiction) has the potential for causing the greatest impact on human society in the shortest time. But whether the signal is decodable, unintelligible, or content-free (unmodulated CW), the mere act of detection creates a way to reply. Whereas the search leading to the detection had to cover significant volumes of phase space for success to occur, anyone knowing the characteristics of the signal (and this will be everyone, because public disclosure is the policy of all SETI programs today) can send an answer, just by transponding back the original signal. Though this requires access to a transmitter and antenna, it is far easier than the original search. Anyone can do it, but should anyone? The SETI Committee of the International Academy of Astronautics worked with the International Institute of Space Law to draft a "Declaration of Principles Concerning Activities Following the Detection of Evidence of Extraterrestrial Intelligence" (39) and more recently, the IAA Position Paper "A Decision Process for Examining the Possibility of Sending Communications to Extraterrestrial Civilizations" (40). The first document, signed by most groups that conduct SETI programs, states that no reply to a detected signal should be sent until a global consensus is achieved, and the second document suggests that the U.N. Committee on the Peaceful Uses of Outer Space may wish to consider (in advance of an actual detection) just how such a consensus might be reached. There are no means of enforcement. The purpose of these documents is simply to highlight the possibility that improving technologies for searching might yield a detection at any moment. The authors suggest that it would be wise to consider any possible reply from Earth while there is no time pressure. These ideas were provided as information to the UN COPOUS in June 2000. Given the press of other immediate items facing that body, substantive deliberations of these issues are not likely to take place any time soon. Should a signal from (or other evidence of) an extraterrestrial technology be discovered in the near future, we can expect that a cacophony of national, or even individual, replies will be broadcast, speaking many ideas in many languages. This is likely to be a far more accurate picture of early twenty-first century Earth, our lack of global governance, and human diversity than are the carefully selected sights and sounds of Earth carried by the Voyager 1 and 2 records (41).

ACKNOWLEDGMENTS

This work was supported by those individual donors (large and small) who enable the SETI Institute to continue its study of the origin and distribution of life in the Universe and its searches for evidence of intelligent life elsewhere. Some of the figures and tables have been reproduced in part or whole from another

review written by the author for the *Annual Reviews of Astronomy and Astrophysics* (16).

BIBLIOGRAPHY

1. DeDuve, C. The chemical origin of life. In *Astronomical and Biochemical Origins and the Search for Life in the Universe*, C.B. Cosmovici, S. Bowyer, and D. Werthimer (eds). Editrice Compositori, Bologna, 1997, pp. 391–399.
2. Joyce, G.F. In *Origins of Life: The Central Concepts*, D.W. Deamer, and G.R. Fleischaker (eds). Jones & Bartlett, Boston, 1994, pp. xi–xii.
3. Dole, S.H. *Habitable Planets for Man*. Blaisdell, New York, 1964.
4. Doyle, L.R. (ed.). *Circumstellar Habitable Zones*. Travis House, Menlo Park, CA, 1996.
5. Walker, J.C.G., P.B. Hays, and J.F. Kastings. *J. Geophys. Res.* 86: 9776–9782 (1981).
6. McKay, D.S., E.K. Gibson, K.L. Thomas-Keprta, H. Vali, C.S. Romanek et al. Search for past life on Mars: possible relic biogenic activity in martian meteorite. *Science* 273: 924–930 (1996).
7. Thomas-Keprta, K.L., et al. Elongated prismatic magnetite crystals in ALH84001-Potential Martian Magnetofossils. *Geochimica et Cosmochimica Acta* 64: 4049–4081 (2000).
8. Pappalardo, R.T., et al. *J. Geophys. Res.* 104: 24015–24055 (1999).
9. Kivelson, M.G., K.K. Hurler, C.T. Russell, M. Volwerk, R.J. Walker, and C. Zimmer. *Science* 289: 1340–1343 (2000).
10. McCord, T., et al. *J. Geophys. Res.* 103: 8603–8626 (1998).
11. Chyba, C.F., and C.B. Phillips. *Proc. Natl. Acad. Sci. USA* 98: 801–804 (2001).
12. Gold, T. *Proc. Natl. Acad. Sci. USA* 89: 6045–6049 (1992).
13. McKee, C.F., and J.H. Taylor. *Astronomy and Astrophysics in the New Millennium*, *Natl. Acad. Sci.* (2001).
14. McKay, C., R.S. Saunders, G. Briggs, M. Carr, D. Crown, et al. Mars landing site selection for sample return. Presented at *29th Annual Lunar and Planetary Sci. Conf.*, Houston, 1998.
15. Malin, M.C., P.R. Christensen, J.L. Bandfield, R.N. Clark, and K.S. Edgett, et al. Detection of crystalline hematite mineralization on Mars by the thermal emission spectrometer: Evidence for near-surface water. *J. Geophys. Res.* 105 (E4): 9623–9642 (2000).
16. Tarter, J. The search for extraterrestrial intelligence (SETI). *Annu. Rev. Astron. Astrophys.* 39: 511–548 (2001).
17. <http://exoplanets.org/>.
18. Doyle, L.R., H.J. Deeg, V.P. Kozhevnikov, B. Oetiker, E.L. Martin, et al. Observational limits on terrestrial-sized inner planets around the cm draconis system using the photometric transit method with a matched-filter algorithm. *Ap. J.* 535 (1): 338–349 (2000).
19. Butler, P. Discovery and study of extrasolar planets: Current radial velocity searches: An analysis. Presented at *24th IAU General Assembly* Manchester, UK, 2000.
20. Cocconi, G., and P. Morrison. Searching for interstellar communications. *Nature* 184: 844–846 (1959).
21. Drake, F.D. Project Ozma. *Phys. Today* 14: 40–42 (1961).
22. Tarter, J. (2001) <http://astro.annualreviews.org/cgi/content/full/39/1/511/DC1>.
23. Oliver, B.M., J. Billingham (eds). *Project Cyclops: a design study of a system for detecting extraterrestrial intelligent life*. NASA Rep. CR 114445, 2nd ed. The SETI League, Inc., SETI Institute, 1996.

24. Rather, J.D. Optical lasers for CETI. In *Bioastronomy – The Next Steps*, G. Marx (ed.), Kluwer Academic, Dordrecht/Boston/London, 1988, pp. 381–388.
25. Dreher, J., and D.K. Cullers. SETI figure of merit. In *Astronomical and Biochemical Origins and the Search for Life in the Universe*, C.B. Cosmovici, S. Bowyer, and D. Werthimer (eds). Editrice Compositori, Bologna, 1997, pp. 711.
26. Ekers, R., K. Cullers, J. Billingham, and L. Scheffer (eds) *SETI 2020: A Roadmap for the Search for Extraterrestrial Intelligence*, SETI Press (Mountain View) 2002.
27. Dixon, R. Argus: A future SETI telescope. In *Progress in the Search for Extraterrestrial Life*, G.S. Shostak (ed.), ASP Conf. Series 74: 355–367, BookCrafters, San Francisco, 1995.
28. Welch, W.J., and J.W. Dreher. The One Hectare Telescope. *Proc. SPIE*, 4015: 8–18 (2000).
29. www.skatelescope.org.
30. Howard, A. et al. An all-sky optical SETI survey. Paper #IAA-00-IAA.9.1.08 presented at the 51st Int. Astronaut. Congr. in Rio de Janeiro, Brazil, 2–6 Oct 2000.
31. Cullers, D.K., and R.P. Stauduhar. Follow-up detection in Project Phoenix. In *Astronomical and Biochemical Origins and the Search for Life in the Universe*, C.B. Cosmovici, S. Bowyer, D. Werthimer (eds). Editrice Compositori, Bologna, 1997, pp. 645–651.
32. Shielded Zone of the Moon is defined in SS22.22–SS22.25 in 1998 edition of ITU Radio Regulations <http://www.itu.int/bredh/rr/>.
33. Table 8.1 in *Evaluating the Biological Potential in Samples Returned from Planetary Satellites and Small Solar System Bodies*, National Academy of Sciences, NRC, Washington, DC, 1998. <http://www.nas.edu/ssb/sssbch8.htm#con>.
34. In *Mars Sample Return: Issues and Recommendations*. National Academy of Sciences, NRC, 1997. <http://books.nap.edu/books/0309057337/html/30.html#pagetop>.
35. <http://sec353.jpl.nasa.gov/apc/Interstellar/03.html>.
36. <http://astp.msfc.nasa.gov/interstelprop.html>.
37. Sagan, C., W.R. Thompson, R. Carlson, D. Gurnett, and C. Hord. A search for life on earth from the galileo spacecraft, *Nature* 365: 715 (1993).
38. Drake, F.D. *Intelligent Life in Space*. Macmillan, New York, 1962.
39. Tarter, J., and M.A. Michaud (eds). *SETI Post-Detection Protocol*, *Acta Astronautica special issue*. 21(2). Pergamon, Oxford/New York, 1990.
40. http://www.iaa.net.org/p_papers/seti.html.
41. Sagan, C., F.D. Drake, J. Lomberg et al. *Murmurs of Earth*. Random House, New York, 1978.
42. Marcy, G., P. Butler, D. Fisher, G. Laughlin, S. Vogt, G. Henry, and D. Pourbaix. A Planet at 5 AU Around 55 Cancri, Submitted to *Ap. J.* (2002).

JILL TARTER
SETI Institute
Mountain View, California

F

FIRST FLIGHT OF MAN IN SPACE

The twentieth century was one of rampant technological development, of unprecedented flourishing of science, a century in which grandiose, fantastic ideas became reality, and when the dreams of many generations of people who lived and now live on Earth came true.

There were a great many ideas, discoveries, projects, and accomplishments in the century just past. But among all the numerous outstanding human achievements in the twentieth century that captured the contemporary imagination, the most thrilling were the launch of the first artificial satellite and, especially, the flight of the planet's first cosmonaut, Yuriy Alekseyevich Gagarin on the Vostok spacecraft.

The launch of the first artificial satellite into orbit around Earth took place on 4 October 1957 at 22:28 Moscow time marking the start of the Space Era in human history. Yuriy Gagarin's spaceflight, launched on 12 April 1961 at 09:07 Moscow time, lasted for 108 minutes, during which the spacecraft carrying the first cosmonaut completed a full orbit around Earth and then landed safely in the scheduled recovery area. This flight marked the start of the Era of Human Presence in Space.

These two events are directly linked; without the first, the second could never have occurred. The first became a momentous milestone in the history of technological progress; the second was the embodiment of a centuries-old dream of humankind, one dreamt both by ordinary people gazing at the starry sky and by the world's great science fiction writers, seers, and creators.

The Prehistory of the First Artificial Earth Satellite Vehicles and Orbital spacecraft

Manned spaceflight is part of the regular process of development in history and reflects the eternal human desire to learn the secrets of nature and to discover

new living environments. The idea of flying to the Moon and the stars arose many centuries ago, when people still did not know either the structure of the solar system; or that Earth is round, rotates on its axis, and revolves around the Sun; or that Moon is its satellite. Indeed they knew very little about the continents of Earth itself. This did not prevent them, though, from creating legends and tales about flights into the sky in which they proposed the most unlikely vehicles for such voyages. In these legends and tales, heroes flew on swans, eagles, flying carpets, winged horses, balloons, and other exotic vehicles. In more recent times, in association with the development of technology, somewhat more scientifically justified projects began to be proposed, for example, giant cannons. The famous science fiction writer Jules Verne sent his hero to the Moon in a missile shot from a cannon that had a barrel approximately 300 meters long. His book is even titled in Russian *To the Moon by Cannon*. As science and technology developed, science fiction writers kept proposing new types of interplanetary craft and materials from which they could be manufactured.

However, the actual means that later made spaceflights possible was only discovered in the nineteenth century. This was the use of rocket engines to propel a spacecraft. Rockets had been used even earlier, for many years before this, mainly for military purposes. As early as the Middle Ages, inventors had begun to propose various ways to use rockets for propelling loads (missiles) across long distances. Rockets were used in war in sieges and taking of forts. The military aspect of rocket use predominated in the majority of inventions associated with them.

In the nineteenth century, the building of rockets had already developed extensively. Rockets were invented that used powder as well as those with liquid-fuel engines. However, the use of rockets for military purposes was still less effective and extensive than the use of artillery tube systems. Nevertheless, enthusiasts, believing in the future of rockets, continued working to improve them.

The scientific principles underlying the use of rockets for spaceflight were developed in Russia through the work of an outstanding self-taught scientist, a schoolteacher from the provincial city of Kaluga, K.E. Tsiolkovskiy. In 1883, Tsiolkovskiy, in his manuscript *Free Space*, advanced the idea that it would be possible to use the principle of reaction propulsion for spaceflight and provided a sketch of a spacecraft that would take humans into space. In 1895, in another of his works, *Dreams of the Earth and the Sky and the Effects of Universal Gravitation*, he gave a rationale for the belief that it was possible to attain the velocity necessary to break away from Earth and demonstrated that it was theoretically possible to build an orbital spacecraft. In 1897, Tsiolkovskiy derived the basic equation for rocket velocity, which is widely known as "Tsiolkovskiy's formula."

Tsiolkovskiy's fundamental scientific work, *Space Exploration Using Reaction Propelled Vehicles*, was published in two parts in 1903 and 1912. In this work, he established the laws of motion for rockets as bodies of variable mass, defined the efficiency coefficient for a rocket, investigated the effects of air resistance on its motion, noted the advantages of rocket engines at high speeds, and provided a sketch of an interplanetary rocket, for which he pointed out the advantages of a liquid propellant. Considering the rocket as the only practical and acceptable means of spaceflight, Tsiolkovskiy did a great deal to define the rational path along which cosmonautics and rocket building had to develop. He published hundreds of scientific works on these problems. It was he who spoke

the prophetic words, "Human beings will not remain on Earth forever; in the pursuit of light and space, they will first timidly penetrate beyond the bounds of the atmosphere and then conquer the space within the solar system." Tsiolkovskiy's ideas were developed very intensively by his students and followers in the 1920s and 1930s. Work on rocket technology was also taking place in other countries. The most significant such work is associated with the names of Robert Goddard (U.S.A.), H. Oberth (Germany), and R. Esnault-Peterie (France). By this time, in the Soviet Union, the efforts of individual rocket building enthusiasts had been combined in the Group for the Study of Reaction Propulsion (GIRD). A special laboratory was also founded and subsequently named the Gas-Dynamics Laboratory (GDL). These organizations played a prominent role in developing various types of rockets, including launch rockets using smokeless powder and liquid rocket engines for aircraft and torpedoes.

In 1933, the Government decided to found the Reaction Propulsion Scientific Research Institute (RNII) to centralize the efforts of the country's rocket builders. The efforts of this institute were responsible for the development of many types of rockets, and liquid-fueled rocket engines for missiles and aircraft. Modified models and new types of rocket that they developed were subsequently used extensively as weapons at the fronts of the Great Patriotic War of 1941–1945 (World War II). Among these, the legendary Katyusha, whose crushing salvos sowed panic and destroyed enemy troops and weapons, was especially popular in the army.

Germany was highly successful in developing and using rocket technology for military purposes during World War II. These successes are associated with the name of Werner von Braun—the Project Director for building the V-1 and V-2 rockets, which were used to bomb London. After the defeat of Germany, technical documentation on these rockets, a few examples of the rockets themselves, and some components found their way to the United States and the Soviet Union. These were used by both countries in the postwar period to build improved military rockets. And von Braun himself went to America and directed work on rocket technology.

As a result of worsening international tensions and the threat that atomic weapons would be used, the Soviet Union resolved to build a powerful intercontinental ballistic missile capable of carrying a warhead many thousands of kilometers. In a short time, this task had been accomplished: on 27 August 1957, the Soviet Union successfully fired one of these rockets. During this period, work on creating multistage intercontinental ballistic missiles was directed by Academician S.P. Korolev, who had been working in the area of rockets since the end of the 1920s. Successful testing of these military rockets showed that spaceflight and the launch of orbital satellites was technically possible.

The First Artificial Earth Satellites—Practical Preparation for Manned Space Flight

On 5 January 1957, Korolev sent to the Government "Proposals for the First Artificial Satellites of the Earth Before the Start of the International Geophysical Year." These proposals were examined and the USSR Government tasked

scientists, leaders of industry, and the military to implement preparations for launching the first artificial satellite of Earth, in full accordance with the program of scientific research for the International Geophysical Year. Preparations for the conquest of space required setting up special scientific institutes and laboratories, industrial enterprises, a cosmodrome, and a network of ground-based tracking stations in the country and training highly qualified work forces. All this had to be done despite the lack of any previous experience anywhere in the world. Nevertheless, the work was completed on an accelerated schedule.

On 4 October 1957, the Soviet Union launched the first artificial satellite in the world into orbit around Earth. The date of that launch entered history as the start of the Space Era.

The first Soviet satellite was a sphere 0.58 meters in diameter and a mass of 83.6 kg. The satellite's two radio transmitters would make available new data about the atmosphere. The successful functioning of the first satellite confirmed the correctness of the theoretical calculations and design solutions underlying construction of the launch vehicle, the satellite itself, and all onboard systems.

After the first launch, other, heavier satellites carrying improved onboard equipment were put into orbit. Meanwhile, the ground-based components of the space infrastructure were undergoing parallel development. The research program implemented in the second satellite included unique experiments with the dog Layka—the first space voyager belonging to a higher animal species. Flights of dogs continued subsequently and made it possible to study the status of a living creature under conditions of weightlessness. A foundation was being laid for the decisive conquest of space through manned spaceflights.

The first U.S. artificial satellite, Explorer 1, was launched into orbit on 1 February 1958. The United States thus became the world's second space power.

In May 1959, the USSR Government adopted two resolutions, "On preparing Humans for Space Flight." Based on these resolutions, the Government conducted a series of experiments involving flights of satellites carrying dogs and human dummies. At the same time, a spacecraft in which a human could be sent into space was being developed. Spacecraft of this type were named Vostok; until the first manned flight, they were launched into space unmanned. Experiments with dogs were not conducted on them.

Orbital spacecraft were launched on

- 15 May 1960—with no living things on board;
- 19 August 1960—with the dogs, Belka and Strelka, on board (recovered safely);
- 1 December 1960—with the dogs Pchelka and Mushka (caught fire on re-entry, the dogs died);
- 22 December 1960—with the dogs, Zhemchuzhina and Zhulka (the capsule failed to go into orbit; the descent module was recovered after 2 days; the dogs survived);
- 9 March 1961—with the dog, Chernuskha, and a human dummy (recovered safely);
- 25 March 1961—with the dog, Zvezdochka, and a human dummy on board (recovered safely).

The missions of these launches were to refine development of the spacecraft design and systems, to conduct biomedical experiments with dogs and other biological subjects, to return the descent modules to Earth, and to implement ejection and parachute landing of a dummy simulating a suited cosmonaut. The generally successful accomplishment of this flight program confirmed spacecraft and onboard system reliability. The prerequisites for manned space flight had been attained in practice.

Preparing the First Cosmonauts for Flight

To prepare humans for the first and subsequent space flights, a team of 20 candidates for flight was selected. Fighter pilots, submariners, rocket builders, automobile racers, and many other young and healthy people dreamed of becoming cosmonaut candidates. The flight surgeons who had been assigned the task of selecting the first spaceflight candidates were well aware that, of members of all of the professions, fighter pilots were most suited to endure the effects of extreme environmental factors. During their training and actual flights, they experienced the effects of hypoxia, increased pressure, G-forces in different directions, ejection, and other factors. During the first phase of selection, it was considered expedient to select the young cosmonaut candidates from among fighter pilots. This idea was fully supported by Chief Spacecraft Designer, S.P. Korolev, who said, "For this enterprise, it would be best to use trained pilots, especially, jet fighter pilots. A fighter pilot is just the generalist that is needed. He flies in the stratosphere in a one-man high-speed aircraft. He is a pilot and navigator and radioman and flight engineer rolled into one...."

Candidate selection began in Air Force units in October 1959. During the initial selection process, documents for 3461 fighter pilots, aged up to 35 years, were examined. A total of 347 men was selected for the initial interview. Based on the results of the interviews and outpatient medical examinations, 206 pilots were selected to undergo further medical selection. These candidates underwent inpatient examination in the Military Clinical Aviation Hospital between October 1959 and March 1960. Of the 206 men, 72 dropped out of their own volition in the course of the examination process. An additional 105 failed to meet medical requirements. Of the 29 pilots who passed all phases of medical examination and met all medical requirements, 20 were finally selected as cosmonaut candidates.

In late 1959, a government decision was made to create within the Air Force a special Center to train candidates for manned space flight. This was done concurrently with selection of cosmonaut candidates. In March 1960, the Cosmonaut Training Center was established. At first, the Center was located in Moscow in the Central Aerodrome. Then, a site outside Moscow was selected. In the summer of 1960, the Cosmonaut Training Center began to operate in Zvezdnyy Gorodok, which had been specially constructed for this purpose.

In early March 1960, the first group of spaceflight candidates, which was still not up to full strength, came to the Center, and on 14 March, the first class in general space training was held for this group. This group was brought up to full strength of 20 by mid-July 1960. Later this first cosmonaut team was named the

Gagarin team. The names of the members of the Gagarin team were I.N. An-ikayev, P.I. Belyayev, V.V. Bondarenko, V.F. Bykovskiy, V.I. Filatyev, Yu.A. Gagarin, V.V. Gorbato, A.Ya. Kartashov, Ye.V. Khrunov, V.M. Komarov, A.A. Leonov, G.G. Nelyubov, A.G. Nikolayev, P.R. Popovich, M.Z. Rafikov, G.S. Shonin, G.S. Titov, V.S. Varlamov, B.V. Volynov, and D.A. Zaikin.

In August 1960, the "Regulation on USSR Cosmonauts" was adopted. This document defined the following Center staff positions: "cosmonaut cadet," "cosmonaut," "cosmonaut instructor," and "senior cosmonaut instructor." The phases of cosmonaut training for spaceflight were defined, as was the list of organizations tasked with conducting this training. Issues concerning the material and social welfare of cosmonauts and their families were also resolved.

A total of only 1 year was allocated to train the first cosmonaut candidates (at that time called cosmonaut cadets) for flight. However, the flight training program was extremely extensive. Training of the first cosmonaut candidates consisted of theoretical classes, training on various simulators, and fieldwork in the design bureau where spacecraft were being built. At the direction of Korolev, classes on rocket technology and celestial mechanics were taught by the most experienced staff members of the design bureau. From their very first days in the program, the cosmonaut candidates were given to understand that classes in these disciplines must form the basis for their future profession.

It was also understood that throughout the entire training period, the future cosmonauts had to be subject to strict and constant medical monitoring, without which it would have been impossible to readjust the demands that were being put on them during training in a timely fashion. The demands on the cosmonauts' health had to be commensurate with the goals and tasks of the future flight.

The instructional and training program for future cosmonauts involved simulating spaceflight factors and conditions. The most complex phases of the flight program used special ground-based facilities, simulators, training devices, mock-ups, and flights on mass-produced or specially modified aircraft. Korolev, who had been a glider pilot, was a vehement advocate of flight and parachute training. He believed that this type of training would polish and refine the cosmonauts' professional skills and would also give them a large infusion of will-power.

In addition to these other factors, the great significance attached to flight and parachute training stemmed from the fact that, on flights on the earliest spacecraft, a cosmonaut had to eject from the spacecraft cabin along with his seat and then, after he had separated from it, descend by parachute.

As a result of this, people from a wide variety of professions went to work to set up and conduct a unified training process for the "special contingent," as the future cosmonauts were then called. The word "cosmonaut" itself was kept secret until the first flight, and it was recommended that it not be used in conversation.

The training program for the first cosmonauts was developed on the basis of their primary mission—to define the limits of human potential to live and work in weightlessness. Gaps in knowledge at that time included the following: Could the cosmonaut "float" in the spacecraft cabin, or did he have to remain strapped to his seat throughout? Wouldn't there be severe psychological disruptions in

space that prevented the cosmonaut from acting responsibly and consciously? There were many other similar questions. For this reason, biomedical training of the student cosmonauts was one of the main types of flight preparation. It was conducted, using specially designed simulators, by leading experts in medicine and psychology, who had a great deal of experience in practical work with pilots on flight qualification and therapeutic and prophylactic problems.

S.P. Korolev considered that the major problem involved in preparing humans for the first spaceflight was ensuring their safety. Safety was considered everyone's responsibility—those who developed space technology, as well as cosmonaut training specialists. They were all focused on solving the following problems:

- ensuring that the cosmonauts were provided with the fullest possible preliminary familiarity (under laboratory conditions and during aircraft flights) with predicted spaceflight conditions, which were reproduced on the training simulators one by one or in combination;
- phased refinement of the flight mission, particularly with regard to using the spacecraft systems;
- development of the cosmonauts' confidence in their own strengths and knowledge and in their readiness to undergo the most severe ordeals.

Here, it would be worthwhile to describe the following very significant phase that occurred during training of the first cosmonauts. When spaceflights with dogs took place as part of the single-orbit flight program, the cosmonauts flew to the launch pad at the Baikonur cosmodrome. After launch, they traveled by aircraft to the reentry module landing site to familiarize them with results of the flight and landing. This was done to teach them about and provide them with direct observation of launch operations, as well as to dispel the excess stress that might be experienced before their upcoming flights.

An important aspect of the training program involved testing the cosmonauts under exposure to prolonged solitude in a "limited space" under various daily work–rest schedules. Initially, the duration of these tests was limited to 15 days. Such tests made it possible to study the individual's psychological reserves.

The professional training provided to the pilots of the earliest spacecraft used a single training simulator. During training on it, the cosmonauts developed the components of the skills and knowledge required to control the onboard systems of the spacecraft. The trainer used a primitive visual simulation system that reproduced the spaceflight environment and incompletely simulated the dynamic operations of spacecraft control. More complex training simulators were only designed and used in preparation for subsequent flights.

To compensate for the inadequacy of the training simulators and other equipment, the facilities of the industrial plants where space technology was developed and tested were used for practical cosmonaut training. The future cosmonauts were frequent visitors to the plants, to the assembly and test building, and to the development stands and facilities. They made themselves at home in a spacecraft cabin mounted on an assembly stand. Lessons at the facilities using actual technology were conducted by the most experienced developers,

designers, and testers of spacecraft and their onboard systems. Frequently, Korolev himself conducted the lessons. He taught the cosmonauts about the design features and possibilities of the first spacecraft and also about the advanced and improved spacecraft of the future.

This combination of all feasible types of cosmonaut training and the means and methods used to provide it ensured that the cosmonauts were trained to a rather high level in the very short period of time allotted (approximately 1 year). This in itself was a unique undertaking, considering that there was no backlog of experience with such matters.

The most important goal in training cosmonauts is to ensure that each one forms an image of the upcoming flight, called by psychologists a conceptual model. During the flight itself, including occurrences of unforeseen or contingency situations, this image functions as a standard of performance and behavior. And, as the cosmonauts reported later, after they had actually flown, the real flight conformed fully to the expectations that they had developed in the course of flight training.

The Manned Spacecraft Vostok

When the concept of the first human spaceflight was being considered, it was proposed that the program should start, not with orbital flight, but with flying a cosmonaut to a great height and returning him ballistically to Earth. This type of flight would have been technically easier, but it would have represented only a timid, momentary type of space exploration. After all, the weightless state lasts only a few minutes during ballistic flight, whereas a single pass in near Earth orbit involves an hour and a half of weightlessness.

Thus, the objective that was set entailed developing a manned satellite capable of functioning in near Earth orbit for several days and then of returning to Earth. All necessary conditions to support normal human vital activity would have to be provided on board this satellite.

This task was daring and difficult. A booster rocket, the satellite itself, and the life support and reentry systems, all possessing extremely high reliability characteristics, were required for the manned flight. Many of these challenges were being confronted for the first time in history. The problem of returning the cosmonaut from orbit to Earth was particularly difficult. The robotic spacecraft of that time were not capable of returning to Earth; yet this was the main prerequisite for a manned craft. Furthermore, not only did a safe landing need to be ensured, but the landing had to occur in a predetermined area. Special safety measures also had to be provided in cases of booster rocket malfunction at launch or during insertion into orbit.

A spacecraft flying at a velocity of 8 km/s (29 thousand km/h) has to be decelerated and landed safely on Earth—at the time, this objective seemed almost fantastic. At that time, aviation was still in the process of mastering supersonic velocities. And here people were talking about 25 times the speed of sound!

When a spacecraft enters the dense layers of the atmosphere at such speeds, a powerful impact wave develops in front of it, and the air in this wave is

transformed into red-hot plasma at a temperature of 6,000–10,000°C. This is higher than the temperature on the Sun's surface. What needed to be done to allow the spacecraft to remain intact throughout this process and to protect the cosmonauts from the heat?

It was proposed to solve this problem by using a design analogous to a walnut. The metal body would be concealed in a "husk," which would burn and evaporate in the course of the descent and prevent the spacecraft itself from getting too hot. A special material was developed for this "husk," but it proved to be relatively heavy. This gave rise to a new problem—how to minimize the total weight of the heat shield. How this problem was solved depended on what method was going to be used to return the spacecraft to Earth. After considering several possible versions, the one selected involved ballistic deceleration followed by parachute-aided descent during the last phase of landing. Calculations showed that this method was simple and technologically feasible.

However, if the entire spacecraft was to descend using this system, the mass of the heat shield it required and the dimensions of the parachute system would exceed reasonable limits. Thus the idea arose of subdividing the spacecraft into the reentry module, which would house the cosmonauts, and the instrument and equipment module, which would contain the retrorocket and propellant tanks, the control system and other technical auxiliary systems needed for orbital functioning. In this spacecraft design, only the reentry module would need heat shielding.

Of all the reentry module shapes considered, the sphere proved the most preferred. It has the minimal surface area for a given volume and thus required the minimum heat shielding mass. It was easy to provide a sphere with stability during reentry by weighting its frontal section. Landing precision was acceptable—plus or minus 300 km. However, the descent of this type of module could only be ballistic.

But could a cosmonaut endure the high deceleration that would inevitably occur during ballistic descent? Calculations showed that at a small atmospheric entry angle of 2–3°, deceleration would not exceed 9–10 G and would continue for only about a minute. According to data from aviation medicine, healthy individuals would be fully capable of enduring such a G-force.

In less than a year and a half, the Soviets succeeded in building and testing a spacecraft system based on a new principle, the most complex that had ever been created at that time. Before the launch of the spacecraft in manned mode, only five test-flights were conducted. But, in essence, all that Soviet cosmonautics had achieved before that time was part of the preparation for this flight. Each test-flight was meticulously analyzed, so that no ambiguities remained. Cosmonaut safety measures were thought out in great detail.

The spacecraft for the first manned spaceflight had been created. From the outside, it looked rather simple: a spherical craft—the cosmonaut cabin—and an instrument and equipment module that was docked with it. The two modules were held together with four metal straps, attached on the "crown" of the reentry module by a pyrotechnic bolt. Before entry into the atmosphere, the bolt exploded, the metal bands were ripped away, and the reentry module continued its movement toward Earth separately from the instrument and equipment module (Fig. 1).

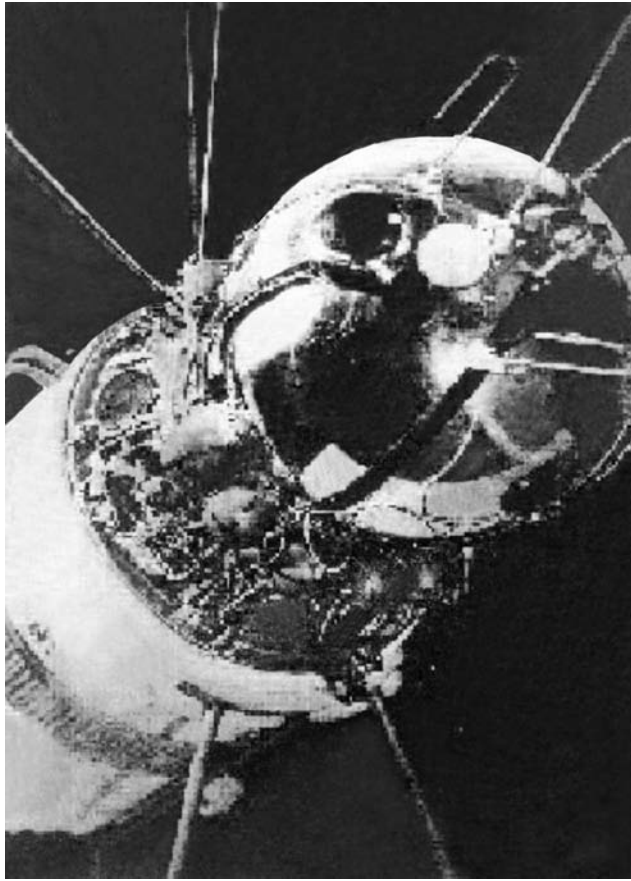


Figure 1. The “Vostok” space vehicle.

The total mass of the spacecraft was 4.73 metric tons; the instrument-equipment module accounted for 2.33 tons of the total. The diameter of the reentry module was 2.3 m, and its mass was 2.4 metric tons. The walls of the reentry module contained heat-resistant windows and fast-opening airtight hatches. In the center of the cabin was an ejection seat for the cosmonaut, who would be wearing a spacesuit. The seat contained the spacesuit ventilation device, the cosmonaut parachute system, and an emergency kit containing everything necessary in case of a landing in an unscheduled region. Thanks to its thermal shielding, the reentry module, though surrounded by a red-hot cloud of plasma created by air resistance, could fly safely through the dense layers of the atmosphere, decreasing its speed from orbital velocity to 180–200 m/sec. At an altitude of 7–8 km, the cosmonaut could eject through the exploded hatch and land by parachute separate from the spacecraft and seat. This system also served as a cosmonaut escape system in case of a malfunction at the start of flight. Special rocket engines installed in the seat could carry it, along with the cosmonaut, from the danger zone and raise him to a height sufficient for safe

parachute landing. The spacecraft was fully automated, but the pilot could take control manually.

The attitude control system, consisting of gyroscopic and optical sensors, logic devices, and microengines running on compressed gas, made it possible to pilot the spacecraft manually. The cosmonaut could control the attitude of the spacecraft in order to turn on the retrorocket. There was a special instrument, called Vzor, that could be used for visual orientation. If the craft was correctly oriented, the cosmonaut could see "Earth passing below" through the central portion of this instrument, that is, the cosmonaut could control the heading. Through the annular mirror, the cosmonaut could see the horizon, which allowed controlling the spacecraft's pitch and roll. The spacecraft was given the name Vostok, which became world famous after its launch with a human on board. The final work on the spacecraft was completed in March 1961, at the same time the first team of cosmonaut candidates completed their training course.

Yuriy Gagarin—Planet Earth's First Cosmonaut

During the training of the first group of 20 men selected as candidates for spaceflight, a smaller six-man subgroup was identified for intensive training for the first flight. The remainder were trained on programs for future flights. The selected subgroup consisted of the following student cosmonauts: V.F. Bykovskiy, Y.A. Gagarin, G.G. Nelyubov, A.G. Nikolayev, P.R. Popovich, and G.S. Titov (Fig. 2).

All six of the candidates selected for the flight were qualified and capable of performing it. But it was necessary to select the best of the best with regard to professional training, health status, psychological traits, and many other qualities. For this reason, each person who had even a slight say in who would be the first cosmonaut carefully observed each of them during the training process as well as in the cosmonaut's free time. Chief Designer Korolev also observed these men carefully.

In the course of training and during various official and unofficial events and meetings, the preference of the majority converged on the same one of the six cosmonauts—Gagarin. He was a born leader. The following lists only a few of the traits he possessed: unshakable faith in the success of the flight, outstanding health, unquenchable optimism, a flexible mind, curiosity, boldness and decisiveness, precision, love of work, restraint, simplicity, modesty, and sociability. Everyone who spoke to him was charmed by his personality.

As had been decided earlier, the decision was made by a State Commission at the Cosmodrome, but it was already clear to many that the first to fly should be Yuriy Gagarin. Events later confirmed that this choice was the right one. Gagarin handled the world's first space flight to perfection, withstood the stresses of fame, and represented his nation with dignity on all the continents of this planet. But in those days of April 1961, to all participants in the project, he was simply the man who would be the first to test a complex technology based on a completely new principle. People had faith in him; people worried about him.

Who was this man, Yuriy Gagarin? He was born to a family of peasants on 9 March 1934 in the town of Gzhatsk in the Smolensk District. He spent his



Figure 2. S.P. Korolev and Yu.A. Gagarin with a group of cosmonauts from the first team.

early years in the village of Klushino not far from Gzhatsk. The Gagarins had four children; Yuriy was the third. In 1941 he started his education at the village school. But the war had already come to the Smolensk area. The village of Klushino, the town of Gzhatsk, and the whole Smolensk area were occupied by the German troops attacking Moscow. Young Gagarin had to live through difficult times. Despite this, he was a good student who displayed a love of learning. In 1949, when he was 15, he decided to leave high school to be able to start helping his parents sooner. His goal was clear—he would work at a plant. For this, he needed to master a trade. He entered a trade school in the town of Lyuberts not far from Moscow. In 1951, he graduated from this school with honors with training as a mold maker and caster; at the same time, he completed an academic school for young workers.

In this same year, he received authorization for further instruction in an industrial technical college in the city of Saratov. In 1955, Yuriy graduated with honors from this technical college and from the aeronautics club where he had also studied while he was a student at the former school. His attraction to flight prevailed, and he entered the First Chkalovskiy Military Aircraft Pilot Academy, from which he graduated in 1957 with a first class qualification. Then, Yuriy Gagarin was sent as a military fighter pilot to serve in the North in one of the naval aviation units of the Northern Fleet. After 2 years, certain young fighter



Figure 3. Yu.A. Gagarin and Mrs. V.U. Gagarin with their children.

pilots were detailed to master a new technology. And in 1960, Yuriy Gagarin found himself in the first team of cosmonauts and then the first candidate for the first spaceflight in history (Fig. 3).

“Let’s Go...”

On 30 March 1960, the nation’s highest leadership received an official memo signed by officials from the Government, Academy of Sciences, War Department, and Industry. The memo stated, “We hereby report that we have performed a large number of scientific research studies and tests on the ground as well as under flight conditions... The upshot of this work to construct a manned orbital spacecraft and a system for returning to Earth, and to provide cosmonaut training is that now we are in a position to implement the first manned spaceflight. Six cosmonauts have been trained to make this flight.” (The memo did not contain the names of these cosmonauts.)

On 3 April 1961, a government resolution was adopted “On the launch of a manned orbital satellite.” This resolution stated: “I approve the proposal for

launch of the Vostok manned orbital satellite with a cosmonaut on board..." On 6 April 1961, the results of the graduation examinations for the first cosmonaut team were ratified, and special certificates were issued to them. A set of instructions was developed, and the cosmonaut's flight mission was signed. This mission was defined as follows: a single-orbit flight around Earth at an altitude of 180–230 km; flight duration—1 hour 30 minutes; flight objective—to verify that it is possible for a human being to survive in space in a specially equipped spacecraft, to flight test the spacecraft and radio communications equipment, and to validate the reliability of the spacecraft and cosmonaut landing devices.

On 10 April 1961, the State Commission approved S.P. Korolev's proposal to implement the first manned space flight in the world on 12 April 1961 on the Vostok spacecraft. The Commission approved Yu.A. Gagarin as the first cosmonaut pilot, and G.S. Titov served as backup pilot.

On 12 April 12, 1961 at 09:07 Moscow time, the first manned spacecraft in the world, Vostok-1, was launched. Its pilot was Major Yuriy Alekseyevich Gagarin, a citizen of the Soviet Union. The flight continued for 108 minutes, during which the spacecraft carrying the cosmonaut made a complete orbit around the Earth. For his successful completion of the world's first space flight Major Yu.A. Gagarin was awarded the title of Hero of the Soviet Union, and also the honorary title of "USSR Cosmonaut-Pilot." The era of human space exploration on piloted spacecraft had dawned.

A government statement regarding the first manned space flight read: "We consider triumphs in space exploration not merely the achievement of our own nation, but of all humanity. We are pleased to place them at the service of all nationalities in the name of the progress, happiness and well-being of all the people on Earth." The world was stunned by the announcement of this event. People on all continents ecstatically greeted the man who had been the first to view our planet from space.

And this is how Gagarin's launch and flight took place. In the last minutes before the launch, everyone who was present felt extraordinary emotion, especially those in the command bunker. The launch commands were listened to with strained attention...

"Fire!"

"Lift-off!"

And then Gagarin's famous words:

"Let's go!"

The first seconds were especially tense for all those in the bunker. The rocket was still relatively low and there remained some risk if the cosmonaut had to eject due to a malfunction, because the rocket could fall relatively close to the ejecting cosmonaut. All listened tensely to the voice of the operator on the loudspeaker "Five... five...five.." This meant that all systems were working normally... The launch went off superbly.

During the flight Gagarin, according to his reports to Earth, monitored the functioning of the instruments, equipment, and spacecraft systems and maintained the requisite constant radio and telegraph contact, observed Earth and the stars, and took water and food, which was also part of his flight program. During the entire flight, the cosmonaut evaluated the effects of weightlessness on his feeling of well-being and his performance capacity.

Everyone impatiently awaited the message from the tracking station at the very south of the country, to which Vostok was to come close. Everyone knew that during reentry from orbit, when the spacecraft passed through the plasma cloud, the transmitter signal would be lost. This disruption of communication was to take place at a predetermined moment. And everyone waited; would it or wouldn't it happen?

"OK, the signal is lost!

And 20 minutes later, a message was received from the Saratov region:

"The cosmonaut is back on Earth. Everything is fine!"

There, on the banks of the Volga, the most thrilling around the world journey of the twentieth century had just been completed (Fig. 4).

Here is how Yuriy Gagarin himself described it: "... I entered the cabin, which smelled like the wind from the fields, they settled me into the seat, and the hatch door closed soundlessly... Now, I could maintain my contact with the outside world, the flight controllers, and my fellow cosmonauts, only by radio... My eyes fell on the clock. The hands showed 9:07 Moscow time. I heard a whistle and a roar that grew louder and louder, and felt the giant craft begin to shake all over and slowly, very slowly break away from the launch pad. The struggle between the rocket and the Earth's gravity began. The G-force began to increase. I felt as if an irresistible force was pushing me harder and harder into my seat. And although it was placed so as to minimize the effect of the enormous G-force on my body, I had trouble moving so much as a hand or foot. But I knew that this would only last a short time as the spacecraft, gathering speed, went into orbit. The G-force kept on increasing. When we got beyond the dense layers of the atmosphere, the nose cone dropped off automatically and flew off somewhere. Through the windows I could see the far off surface of the Earth. At that moment Vostok

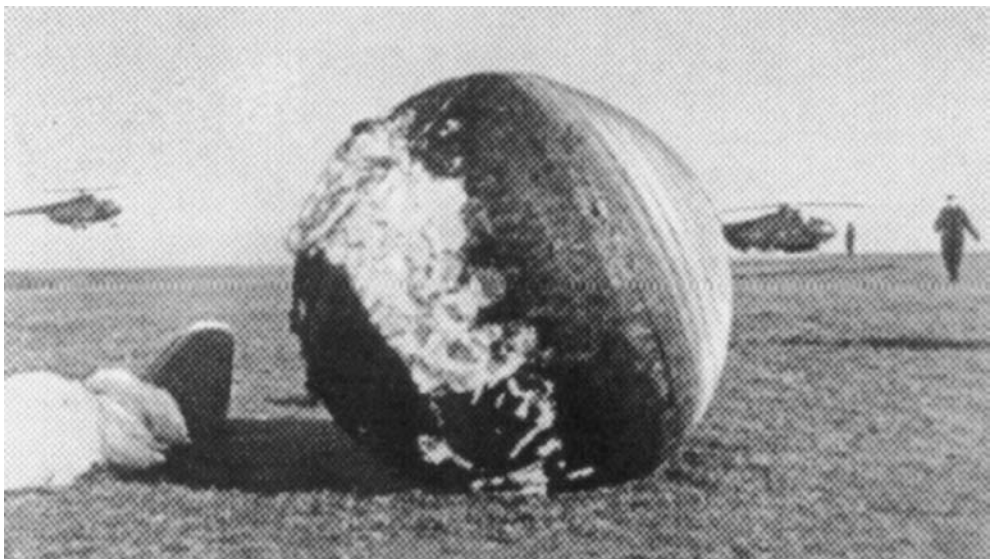


Figure 4. Reentry module of the "Vostok" space vehicle after landing.

was flying over a wide Siberian river. I could distinctly see its islands and shores in the sunlight. The spacecraft went into orbit—on the broad highway of space. Weightlessness started... At first this sensation was strange, but I soon got used to it, adapted and continued to perform the assigned flight program. Vostok was traveling at a speed close to 28,000 km/h. Such a speed is difficult to imagine on the Earth... Then the final phase of the flight approached. This phase—return to Earth—is perhaps even more critical than insertion into orbit and orbiting. I began to prepare myself for it. What awaited me was a shift from weightlessness to a new, perhaps even stronger G-force, and the colossal heating of the exterior spacecraft shell as it entered the dense layers of the atmosphere. Until now everything on the space flight had occurred approximately the way we had worked it out during our training on Earth. But what would happen during the final, culminating phase of the flight? Were all the systems operating normally? Was there perhaps some unforeseen danger awaiting me? Automatic mechanisms are all very well, but I checked the spacecraft's location and got ready to take control into my own hands... At 10:25 the retrorocket fired automatically. It worked beautifully, exactly on time. The flight altitude began to drop. Convinced now that the spacecraft would reach Earth safely, I got ready for the landing. Ten thousand meters... Nine thousand... Eight... Seven... Below, the ribbon of the Volga shimmered. I immediately recognized the great Russian river and its banks..." (Fig. 5).

The excitement produced by Gagarin's flight was triumphal. People throughout the world were thrilled by his exploit. The newspapers of all nations featured articles about this epochal event as their main story under headlines such as: "108 Minutes that Shook the World," "A Fairy Tale Comes True," "Columbus of the Universe," "A New Era," etc. April 12 became the Day of Cosmonautics, a day dedicated to aviators and the conquerors of space, which is celebrated worldwide.

During those days, specialists in public opinion concluded that there was no more famous man in all the world than Gagarin. He was greeted by all the people of Earth, in all languages, as the hero of the twentieth century. He was honored by kings and presidents. He received the highest awards, and babies were named after him. Towns and villages, avenues and squares, seagoing vessels and educational institutions now bear the name Gagarin. There is a "Yuriy Gagarin" crater on the Moon. There are monuments to Gagarin in many cities of the world (Fig. 6).

Gagarin endured the ordeal of fame with dignity. He did not imagine that he was a superman. He remained the man he had been before: simple and human, with the same open-hearted character and the same warm smile. When he was asked whether his fame interfered with his life, he answered, "There is 'fame' and then there is 'Fame.'" The Fame that you feel should be written with a capital letter never was and never will be something that belongs only to you. It belongs primarily to the nation that brought you up and educated you. And such Fame doesn't turn your head. Such Fame makes you demand more of yourself, it is difficult, but reliable."

Surrounded by attention, Gagarin continued to seek new knowledge and new accomplishments. He continued training for new spaceflights and was the backup for cosmonaut V.M. Komarev, who, in April 1967, tested the first Soyuz



Figure 5. Yuriy Gagarin reports to the national leadership on the successful completion of his spaceflight.

spacecraft, which was based on a new principle (the cosmonaut died on landing as a result of failure of the spacecraft's parachute system). Never discontinuing his professional training, Gagarin studied in the Professor N.Ye. Zhukovskiy Academy of Military Aircraft Engineering and on 17 February 1968 successfully defended his thesis. He was awarded a diploma for graduating from this renowned educational institution.

Only 1 month and 10 days after this, on 27 March 1968, Yuriy Gagarin was killed, along with his instructor, while performing a training flight on a training aircraft. This tragedy was mourned by the entire nation. However, Gagarin's name and his exploit will remain in humanity's memory for centuries. The enterprise that Gagarin began continues. After the first Vostok, new spacecraft went into space. American astronauts performed two suborbital (ballistic) flights, and then the United States went on to orbital flights.



Figure 6. The first cosmonaut of planet Earth, Yuriy Gagarin. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

There were six flights of Vostok series spacecraft in the Soviet Union. This vessel was replaced by the multiman Voskhod spacecraft, on which two crews, of three and two men, completed flights. Then Soyuz went into orbit in space. After a series of eight group and one-man flights on spacecraft of this type, they became the main transport vehicles for flights to the manned orbital stations, which began to be launched into space in 1971. Parallel Soyuz spacecraft performed autonomous targeted flights. The first Apollo–Soyuz international experimental program of 1975 used this spacecraft. The total number of flights of Soyuz (and subsequent Soyuz-T and Soyuz TM) series spacecraft, counting supply flights to orbital stations, performed by the Soviet Union and then Russia are Soyuz: 40; Soyuz T: 15; and Soyuz TM: 32.

Extensive programs of scientific studies and experiments were conducted on the Salyut-1, Salyut-3, and Salyut-4 orbital stations as part of a national

research program for the study of space. The research performed on Salyut-6 and Salyut-7 formed part of international collaboration (the Interkosmos program) as well as national programs. Visiting crews, including citizens of various countries, worked on these stations jointly with the prime crews. The most famous orbital station was the multimodular Mir, which functioned in orbit for 15 years (from 1986 to 2001). An extensive program of international collaboration (the Mir-Shuttle, Mir-NASA, and Euromir subprograms and others) took place on this station.

Experience with long duration flights on Mir and the construction and technological design solutions validated on it were used extensively in creating of the International Space Station. Experience gleaned in the United States during implementation of the national manned Mercury, Gemini, Apollo, Skylab, and Space Shuttle programs was also incorporated here. Several international space crews have already gone into orbit on the International Space Station using Russian transport vehicles of the Soyuz TM series and the U.S. recoverable Space Shuttles (Altantis, Endeavor, Discoverer). Information about the general nature of flights on orbital stations of the Salyut and Mir series and also the International Space Station is presented in the article by Yu.P. Semyonov and L. Gorshkov on Russian Space Stations elsewhere in this encyclopedia.

Space has become an arena for international collaboration for the benefit of people throughout the world. The road to this achievement was laid by Yuriy Gagarin, the first cosmonaut of planet Earth.

PETR I. KLIMUK
G.I. VOROBYOV
Yu.A. Gagarin Cosmonaut
Training Center
Russia

G

GLOBAL NAVIGATION SATELLITE SYSTEM

The Global Navigation Satellite System (Glonass) was designed to provide global real-time determination of the position and velocity of an unlimited number of moving objects at any point on the Earth's surface, in the air, and in space. This system was developed in response to an order from the USSR (Russian Federation) Ministry of Defense. By instruction of the Russian Federation's President, the Glonass system was put into operation on 24 September, 1993. A 7 March, 1995 decree of the RF Government put the system at the disposal of the world community in standard mode intended for civil, commercial, and scientific use without a user fee.

A RF Presidential decree of 18 February, 1999 designated the Glonass as a dual-purpose (military and civilian) space navigation system. The decree designated the Federal executive agencies responsible for its use, maintenance, and development as the Ministry of Defense and The Russian Aviation and Space Agency. Issues relating to developing and implementing the systems were to be coordinated by the Internavigatsiya Interagency Commission and an interagency task force set up in accordance with the Russian Federation government decree of 29 March, 1999.

Users are informed of system status by the Coordinating Scientific Information Center of the RF Ministry of Defense, as well as by the Information Analytic Center for the positional and time support of the Russian Aviation and Space Agency's flight control center.

The system makes it possible to obtain highly accurate navigation location information with a maximum error less than or equal to 50–70 meters for position and 15 cm per second for velocity from any point on the globe or in near-Earth space. Information is received in real time after a two to three minute pause, when the user's navigational system has just been turned on and

continuously thereafter. At the same time, the system provides the capability to link the user's timescale with the State unified timescale with an error of no more than one microsecond. If the user's equipment can implement special methods for processing navigational information or differential operating modes are used, the accuracy with which the user's position is determined may be improved significantly. Tests have shown that the maximum error in such cases does not exceed a few meters.

Concepts Involved in Positioning

Users of the system, who are in the continuous radio-navigation field created by the space navigation system, navigate by determining their position and velocity at any moment relative to the navigation satellites. Radio-navigational signals are continuously emitted by each navigation satellite in an orbital constellation. These signals are received by the user's equipment, without interrogation, and are used to determine the relative pseudorange and pseudovelocity (rate of change in pseudorange). Pseudorange is determined on the basis of the time delay it takes the signal to go from the navigation satellite to the user (on the basis of phase shift of the signal obtained from the satellite's relatively stable signal and the user equipment's generator frequency and time store) and pseudovelocity on the basis of the Doppler shift of the received signal frequency. Measured values of pseudorange and pseudovelocity can be used to solve the navigational problem and determine the user's position and speed. Because atomic standard clocks are used on each navigation satellite, the system synchronizes the radio-navigational signals emitted by the satellite with the ground-based highly stable atomic standard that functions as the system standard for frequency and time. Use of the ground-based frequency and time standard and the highly stable onboard frequency standard systems time is kept synchronized for each satellite.

The user's intrinsic time differs from the satellite system time. For this reason, initially, during receipt of signals from the satellites, the frequency of the user's generator is synchronized with the frequency of the signals received. The discrepancy between systems time and the user's clock time is determined at the same time as the coordinates and the components of the user's velocity vector when the navigational problem is solved.

Solution of the navigational problem requires, along with the measured values of pseudorange and pseudovelocity, the use of the satellite ephemeris. At each moment, they determine the satellite's radius vector and velocity vector components with high accuracy. The ephemerides, along with the frequency-time corrections of the onboard generator of each navigation satellite relative to the system frequency and time standard, are transmitted in the radio-navigation signal and received by users.

To solve the navigational problem (i.e., to determine the user position and velocity), the pseudorange and pseudovelocity to four or more navigation satellites must be measured. Using the measured values of pseudorange and pseudovelocity, with known satellite position and velocity (from the ephemeris information), the navigational equipment estimates values of the measured

parameters. As a result of comparing the measured and estimated values of the parameters, the navigational equipment determines the location, velocity, and displacement of its own timescale relative to the Glonass time-scale system.

The accuracy of the user's measurement of position and velocity are influenced by the error associated with the characteristics of the parameters measured by the user's receivers and the mathematical methods used to process them, that is, to solve the navigational problem.

The first group of errors results mainly from errors in the ephemerides and synchronization of each satellite's onboard generator relative to the central synchronizer, which is the ground-based system standard for frequency and time. These errors occur as a result of imperfections in the procedures for comparing the frequency of the onboard and ground-based generators and maintenance of the onboard timescale. For example, a time shift of 10 nanoseconds can lead to an error of 3 meters in measuring pseudorange. The ground-based measurement complex corrects the onboard timescale so that the standard deviation of the shift relative to the timescale of the central synchronizer does not exceed 10 nanoseconds.

The error in the ephemerides is a result of the error in measuring the parameters of the satellite orbit based on trajectory measurements conducted by ground-based command tracking systems in an interrogating mode. The error in the ephemerides is a consequence of measurement errors made by the command tracking system, errors in their relative position in space, position relative to the common ground-based ellipsoid, and the appropriateness of the model of the forces acting on the satellite in flight, which are factored in when the parameters of the orbit and prediction of movement are estimated.

The second group of errors results from errors in the model of radio-signal propagation in the troposphere and ionosphere along the navigation satellite-user path. The magnitudes of tropospheric and ionospheric refraction change as a function of the length of the path that the signals must traverse and the state of the ionosphere and troposphere along this path. To diminish the effect of these errors, the user's equipment uses navigational signals from those satellites that are above the plane of the user's horizon by $5\text{--}10^\circ$. This curtails the length of the path that the signal must traverse through the troposphere and ionosphere. A method that involves measuring the pseudorange and Doppler shift at two frequencies is used to compensate for the error due to ionospheric refraction. This is the so-called two-frequency method. The remaining error using the two frequency method is proportional to $[\sin(\gamma)]^{-1}$. This constitutes 1–2 m at an angle $\gamma = 10^\circ$ between the plane of the user's horizon and the flight direction to the navigation satellite. This method of computing ionospheric refraction is the most accurate; however, its use requires highly precise measurements at two frequencies, which increases demands on the user's equipment and leads to a significant increase in the error component from radio interference.

The current error rate in measuring pseudorange is the result of re-reflected signals from Earth, the sea, and other nearby surfaces. These errors, in many respects, are a function of the spatial relationship between the navigation satellite, the receiving antenna, and the reflecting surfaces. The range of the errors is 0.5–2 m in the best case and reaches 100 m in the worst, for example, under urban conditions with high buildings.

The third group of errors is associated with the user's equipment and is linked to errors in marking the moment that the radio signal is received. The most important contribution comes from the noise and dynamic error in the tracking circuit due to the lag in the bending and carrier signals. Typical error due to noise and quantization of the signal ranges approximately from 0.2–1 m.

The measurement accuracy of user location and velocity, which is affected by the error sources listed, depends substantially on the configuration of the constellation of navigation satellites used to solve the navigational problem. An important condition for achieving accuracy in navigational measures is the relative spatial location of the satellite constellation and the user. This is the basis for the concept of geometric dilution of precision, which is a measure of the diminished accuracy of navigational measurements resulting from the specific features of the relative spatial positions of the satellite and the user. In the ideal case, the most accurate positioning for a ground-based user is attained when the optimal constellation is used, that is, when the user is in the center of a regular tetrahedron. Then, the value of the geometric dilution of precision is equal to 1.5.

The Glonass system consists of three segments: the space segment comprising the constellation of navigation satellites; the control segment—the ground based control complex; and the segment represented by the user's navigational equipment.

The Space Segment

The configurations of orbital groupings of navigation satellites are selected so that the user, at any point on the Earth's surface and at any moment, may work with a constellation close to the appropriate optimum that permits navigational measurements with the prescribed characteristics.

In accordance with this requirement, a fully deployed constellation of the Glonass space navigation system includes 24 navigation satellites located in three orbital planes, separated by 120° of longitude from each other. There are eight satellites in circular orbits in each orbital plane. The altitude of the satellites' circular orbits is about 19,132 kilometers, a satellite's period of rotation around Earth is 11 hours 15 minutes and 43 seconds. The plane of the orbit is inclined by 64.8° to the plane of the equator. The satellites are regularly spaced at 45° from each other in an orbital plane. The phase shift of a satellite's position in one plane with respect to the position of a satellite in another plane is 15° . This orbital configuration makes it possible for a user, at any point on the surface of Earth, in the air, or in space up to an altitude of 2000 km, to receive radio-navigational signals from between 5 and 11 navigation satellites at the same time, depending on the area where the user is located, and to process the measurements from all of the satellites or to select a grouping closely approximating the optimum.

The value of the geometric dilution of precision K_g , associated with an orbital constellation and the probability P_N of seeing a given number of satellites by users on the surface of Earth are shown in Table 1.

The Glonass satellite period of revolution around Earth differs from the geosynchronous period. Because of this difference, the path of each satellite in

Table 1. The Geometric Dilution of Precision K_g and the Probability P_N of Seeing a Given Number of Satellites

Parameters	Number of visible satellites					
	4	5	6	7	8	9
P_N	1	1	1	1	0.91	0.58
K_g	2.45	2.16	2.05	1.91	1.86	1.79

the constellation on the surface of Earth is displaced approximately 21° from the equator every 24 hours. The interval of repetition of a satellite's path in its zone of radio-visibility to ground-based facilities is 17 passes (7 days, 23 hours, 27 minutes and 28 seconds). Because of this, the contribution of resonant disturbances resulting from the eccentricity of Earth's gravitational field is diminished by a factor of 8–10 compared to the disturbance of orbits that have a geosynchronous period of revolution. This leads to decreased displacement of each satellite in an orbital plane relative to a given point. The orbital structure in this case is deformed less over time, is more stable, and requires less energy to maintain a given configuration in the orbital grouping.

Satellites that have system numbers 1–8 are located in the first orbital plane, those that have numbers 9–16 in the second plane, and those numbered 17–24 in the third plane. The system numbers of the satellites in an orbital plane increase in the direction opposite to the satellites' motions. The nominal values of absolute longitude of the ascending node of orbital planes, which is fixed at 0:00 Moscow standard time on 1 January, 1983 is equal to $251.4^\circ + 120^\circ (k-1)$, where k is the number of the orbital plane. The nominal distance between satellites in one plane is 45° latitude.

The argument of latitude of satellite No. 9 (the first satellite in the second orbital plane) is 15° greater than the argument of latitude of satellite No. 1. The latitude of satellite No. 17 (the first satellite in the third orbital group) is 30° greater than the latitude of satellite No. 1.

The Navigation Satellite

The satellite is the main component of the Glonass system. The satellite consists of a cylindrical pressurized housing containing the instrument module, the solar array panels with controls, the framework of the antenna-feeder devices, the instruments of the attitude control system, the modules of the propulsion system, and the louvers of the thermal regulating system with its controls. Optical angular reflectors are mounted on the satellite. These are used for calibrating the measurement systems by measuring distance in the optical range obtained by ground-based quanto-optical systems (Fig. 1).

The satellite's equipment provides high-quality navigational measurements and emits highly stable navigational radio-signals of two types—standard accuracy (SA) and high accuracy (HA) in the decimeter bandwidth; receives, stores, generates, and transmits navigational information; generates, codes, stores, and

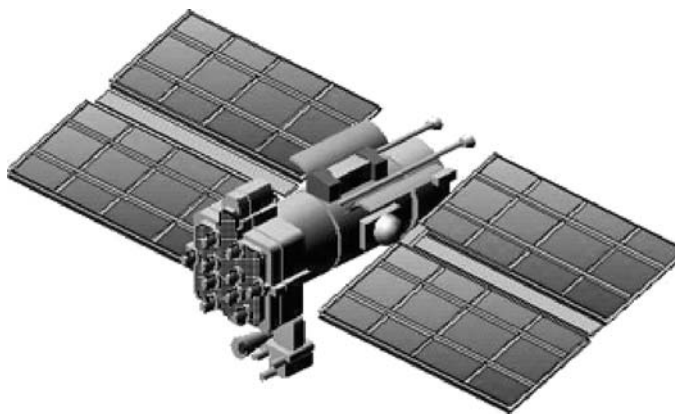


Figure 1. Glonass navigation satellite. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

transmits time signals; receives and processes codes for correcting and phasing onboard timescales; retransmits and emits signals for radio-monitoring of the satellite orbit and determines adjustments to the onboard timescale; analyzes the status of onboard equipment and generates control commands; receives, confirms receipt, decodes, and processes one-time commands; receives and processes programs controlling satellite modes of operation in orbit; generates and transmits signals “Calling Ground Control” when there is a breakdown or important monitored parameters exceed the limits of the norm; and generates telemetry data on the status of onboard equipment and transmits them to the program control office.

The functions listed are performed by the satellite onboard navigational transmitter; the onboard timescale generator; the onboard control system; the attitude control and stabilization system; the systems responsible for orbital correction, electric power supply, and thermal regulation; the onboard fuel and environmental maintenance devices; the structural components; and the cable network. This equipment is continually being improved to lengthen the active service life of the satellite, increase the accuracy of onboard frequency standards, increase the duration of autonomous operation, increase the level of automation of control, and expand functional capabilities. The most important elements of the onboard equipment have backups.

The onboard navigational transmitter generates and emits highly stable radio-navigational signals in two frequency bands, L1 and L2. The Glonass system uses frequency separation of navigational satellite radio signals in both bands. Each satellite transmits navigational radio signals at its own particular frequency in the L1 and L2 bands. Satellites at opposite points of the orbital plane (antipodal satellites) may transmit navigational radio signals at the same frequencies without causing interference for users on the ground.

Each satellite emits radio-navigational signals in the direction of Earth using a transmission antenna. The operating portion of its directional pattern is $2\varphi_0 = 38^\circ$ wide. The axis of the pattern is aimed at the center of Earth. Thus, each satellite’s radio signal covers Earth’s disk up to an altitude of 2000 km.

The navigation satellite transmits navigational radio signals of two types, standard and high precision along the radio links of the L1 and L2 frequency bands. The standard precision signal with a cycle frequency of 0.511 MHz is intended for use by Russian and foreign citizens. The high precision signal with a cycle frequency of 5.11 MHz is modulated by a special code and is not recommended for use without coordination with the RF Ministry of Defense. The standard signal is available to all users who are equipped with the appropriate user equipment and have Glonass satellites in their visibility zone. The characteristics of the standard navigational signal are not intentionally degraded. In subsequent versions of the navigational satellites (Glonass-M), plans call for offering consumers a standard precision code in the L2 band.

The nominal values of the carrier frequencies for satellite navigational radio signals for each band are defined by the following expressions:

$$f_{k1} = f_{01} + k \times \Delta f_1, \quad f_{k2} = f_{02} + k \times \Delta f_2, \quad (1)$$

where $k = 0, 1 \dots 24$ —the number of the carrier frequency designators for the radio signal and

$$f_{01} = 1602 \text{ MHz}, \quad \Delta f_1 = 562.5 \text{ kHz}, \quad f_{02} = 1246 \text{ MHz}, \quad \text{and} \quad \Delta f_2 = 437.5 \text{ kHz}$$

are the carrier frequencies for L1 and L2, respectively. For each satellite, the effective frequencies L1 and L2 are coherent and are generated from the common frequency standard. The nominal value of the frequency of this standard, from the standpoint of an observer on Earth's surface, is equal to 5.0 MHz. The interface control document stipulates a stage by stage change in the frequency range of the Glonass system in the direction of decreasing the number of designators to 12. The distribution of frequencies for the L1 band, in accordance with numbers of designators of radio signal carrier frequencies in the program phases, is shown in Fig. 2.

The navigational signal in the L1 band contains the range code, the onboard timescale and navigational data (ephemeris information, correction of the time, frequency and phase of the onboard frequency standard). The navigational signal in the L2 band contains the range code. It is used to diminish the effect of ionospheric refraction through the two-frequency method on the accuracy of navigational parameter measurements and is intended for special users.

The onboard clock assures continuous transmission of highly stable synchronized frequencies in the navigation satellite system and generation, storage, and transmission of the onboard timescale. It generates signals of standard precision and networks of synchronized frequency impulses. The main component of the clock is the atomic frequency standard.

All navigation satellites are equipped with cesium frequency standards. The precision of mutual synchronization of the onboard timescale is 20 nanoseconds (standard deviation). The basis for generation of the system timescale is the hydrogen standard of frequency of the central system synchronizer, whose diurnal instability is 5×10^{-14} . The discrepancy between the system timescale and the scale of the State standard for the Universal Time Code UTC (SU) should not exceed 1 ms.

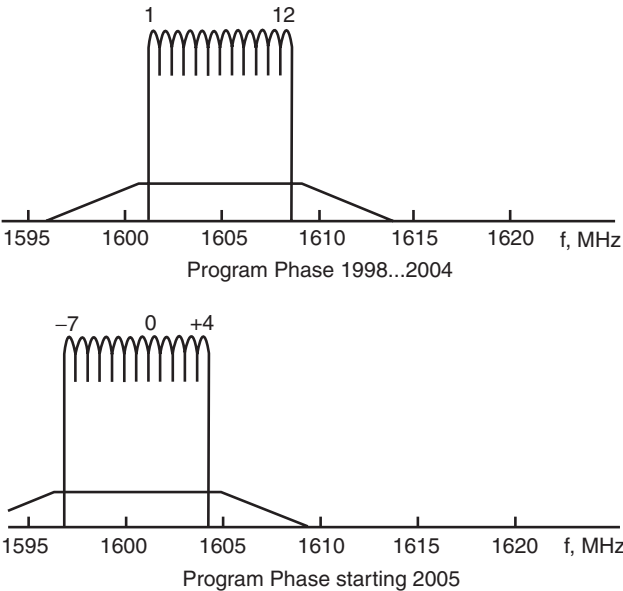


Figure 2. Distribution of frequencies in the L1 band by designator numbers.

The timescale of each Glonass satellite is periodically checked against the timescale of the central synchronizer. Corrections of the timescale of each satellite with respect to the system timescale are computed at the System Control Center and are transmitted to each satellite twice a day. The error involved in checking the satellite timescale against the system timescale does not exceed 10 nanoseconds when it is measured.

The system timescale is corrected at the same time as the correction of the whole number of seconds in the UTC scale, which is regularly conducted by the Universal Time Service. This correction occurs at 00 hours, 00 minutes, 00 seconds during the night between June 30 and July 1 or from December 31 to January 1. The users of the Glonass system are informed ahead of time of the regular implementation of the second correction of the system timescale. Thus, between Glonass system time and UTC(SU), there is no discrepancy by a whole number of seconds. However, between Glonass system time and UTC (SU), there is a constant shift of three hours, corresponding to the system time at the Ground Control Complex.

The onboard control system includes the onboard command equipment, the onboard computer system, the onboard telemetry system and the control module. The onboard computer system stores and processes navigational information and generates and transmits navigational images.

The attitude control and stabilization system is used to stabilize the satellite and provide its initial orientation to the Sun and Earth, orient the satellite's vertical axis to the center of Earth and the solar array to the Sun, orient the thrust vector of the engines to maintain the satellite in its position, and provide orbital correction. The satellites use an active three-axis system for

attitude control and stabilization that has a control flywheel and a jet momentum unloading system.

Launch Facilities

A Proton-K launch vehicle with a 11C861 upper stage (the Proton-KM launcher with the Briz-M upper stage developed at the Khrunichev Space Center) is used to insert the navigation satellites into orbit. These launches take place from the Baikonur spaceport. Groups of three satellites at a time are inserted into the given plane directly into one of the points of the orbit that needs to be filled. At this point, all three satellites separate. One of the three remains at that point. Its orbit is corrected to compensate for possible errors in insertion. The other two satellites are then moved to their assigned points in the orbital plane.

Positioning of a satellite at its assigned point in the orbital plane occurs in stages. These include measurement of insertion orbit parameters and generation of a positioning program. In accordance with the program and using the onboard orbital correction engines, the satellite is decelerated or accelerated. After acceleration (deceleration) a phase of unpowered flight begins, during which the satellite is moved to its assigned point. Once it approaches this point, the correction engine is again turned on to stop the motion of the satellite in the plane and to adjust the orbital parameters in accordance with the assigned position. After each satellite has been positioned at its assigned point, the orbital parameters are precisely determined and included in the system. The duration of the spacecraft positioning at its assigned point can vary and may take from one week to one month, depending on the distance between the insertion point and the assigned point.

Possible variants for filling out an orbital constellation with single satellites using Rus' type launch vehicles, such as converted launchers of the Rokot type, and the newly developed Angara type launch vehicle, are being considered.

Ground Control Complex

The Ground Control Complex tracks the satellites and provides them with the information they need to maintain the operation of onboard equipment and to generate radio-navigational signals and navigational messages. It also monitors the technical status of the satellites and the accuracy of the navigation time field generated by the system.

The main tasks of the ground control and monitoring complex are maintaining the necessary trajectory and other measurements for computing the predicted values of the ephemerides and other auxiliary information, sending ephemeris information and frequency and time corrections to the satellite, monitoring the navigational and time field, generating a unified system timescale, telemetry monitoring, transmitting command and program information, supporting orbital correction, and positioning the satellite in its assigned orbit.

The complex consists of the System Control Center, a network of monitoring and measuring stations located throughout the territory of Russia. The monitoring stations are centers for tracking satellite signals and collecting the

information needed to determine the ephemerides, time corrections relative to the system timescale, and frequency corrections relative to the system standard of frequency and time. The System Control Center collects and processes data to determine the predicted ephemerides and model parameters for onboard clocks, as well as other data to send to the onboard satellite equipment. Communications between the ground command and control complex and the satellites uses the radio channels of the Command Tracking Stations.

The Ground Control Complex contains the following interconnected stationary spatially dispersed components: the Systems Control Center, the Central Synchronizer, the Command Tracking Stations, the Phase Monitoring System, the Quanto-Optical Stations, and the equipment for monitoring the navigational field. All components of the Ground Control Complex are located on Russian territory close to the following geographical points: Krasnoznamensk in the Moscow Oblast (System Control Center), St. Petersburg (Command Tracking Station 9), Shchelkovo in the Moscow Oblast (Command Tracking Station, Phase Control System, Central Synchronizer, Navigational Field Monitoring Equipment), Yeniseysk (Command Tracking Station 4), Ulan-Ude (Command Tracking Station 13), Komsomolsk on the Amur (Command Tracking Station 20, Quanto-Optical Station, Navigational Field Monitoring Equipment).

The Ground Control Complex controls the flight and operation of the satellite's onboard systems by generating control commands and transmitting them to the satellites; making trajectory measurements for determining and predicting orbital parameters; and computing ephemerides for all satellites; measuring time and determining the divergence of all satellite onboard timescales from the system timescale, synchronizing the onboard timescale of each satellite with the central synchronizer time scale and that of the Universal Time Service; generating files of auxiliary information, navigational messages containing the predicted satellite ephemerides, and of the Glonass system almanac; correcting timescales of each satellite and other data for developing a navigation framework; transmitting the files of auxiliary information to each satellite and monitoring their receipt; monitoring the operation of satellite equipment using the telemetry channels and diagnosis of their status; monitoring of information in the navigational messages of the satellites and receipt of ground control call signals; monitoring the characteristics of the navigational field; determining the phase shift of the distance measurement navigational signal with respect to the phase of the central synchronizer; planning the work of all the technical equipment for the Ground Control Complex; and automated processing and exchange of data among all Ground Control Complex components.

Ephemeris Support

Satellite motion parameters are measured and predicted in the ballistic sector for ephemeris-time support based on the results of trajectory measurements obtained by the Command Tracking Center in the interrogating mode. Ephemeris is measured in routine control operations, during which preliminary processing of trajectory measurement occurs and motion parameters are measured; ephemeris information is computed, and the system almanac is generated; a

posteriori estimate of the precision of ephemeris information takes place; and the parameters of Earth's rotation are processed and predicted.

These tasks start with the processing of trajectory measurements obtained during the past day. After preliminary processing of newly received measurements, the parameters of satellite motion are determined with more precision. During precise determination of motion parameters, the constant components of measurement error are determined for each measurement station. Trajectory information obtained during an 8-day interval of motion is used to solve this problem. The problem is solved by the least squares method, and the full sample of measurements during the 8-day satellite motion interval is used.

The more precisely determined initial conditions determining satellite motion parameters are used for computing ephemeris information and the system almanac. This data is in turn used to generate auxiliary information, which is transmitted to the Command Tracking Station for relay to the onboard satellite equipment.

A system of differential equations for satellite motion, solved numerically, is used to define orbital parameters and compute ephemerides precisely. The system of differential equations takes account of the disruptive factor caused by Earth's gravity, including anomalies of Earth's gravitational field to the eighth order of magnitude inclusively; the gravitational field of the Moon and Sun; the effect of lunar-solar tidal disturbances on Earth; light pressure factoring in the specific features of changes in the reflecting properties and the magnitude of the satellite surface turned toward the light.

During computation, the precision characteristics of the ephemerides, which are computed daily for each satellite, are evaluated. Considering the more precise ephemerides determined during the current day, the standard parameters (averaged over the measurement interval) are computed, and the maximum deviation of the evaluated ephemerides from the standard area is derived. The vector of maximum deviations is entered into the database and used to compute sampled evaluations of the precision of ephemeris support for a given time interval for individual satellites or the whole system. At the end of every month, the quality of the measurement system's operation is evaluated. The standard deviations of the satellite ephemerides in the orbital system of coordinates for daily prediction are shown in Table 2.

During the computation of the ephemerides, the coordinates of Earth's pole and changes in the duration of a day are also derived. Specially developed methods make it possible to determine universal time as well. The accuracy of the results obtained are evaluated at the level of 15–20 cm for the pole

Table 2. Error in Navigation Satellite Ephemeris

Error component	Standard error of ephemeris	
	Location, m	Velocity, cm/s
Along the orbit	20	0.05
Across the orbit	10	0.1
Along the radius-vector	5	0.3

coordinates, 0.5 ms for the duration of Earth days, and 1 ms for universal time. Glonass has been regularly deriving parameters of Earth's rotation based on data from satellite observations in real time since 1984.

The derivation of these data is methodologically and administratively associated with the information support hardware of the Glonass systems, which determines the reliability and precision (acceptable for practical purposes) of the values obtained for Earth's rotational parameters. However, the regional location of the Command Tracking Stations exclusively on Russian territory and the features of the orbital structure of the system introduce certain issues into the technology for deriving Earth's rotational parameters.

Derivation of these parameters occurs during the technical phase of controlling the satellites, which involves daily computation of the satellites' orbits and Earth's rotational parameters based on observational data for each satellite during the previous 8 days. Each such derivation generates three values for Earth's rotational parameters—two coordinates for the pole, X_p and Y_p , and the rate of rotation. The current values of the polar coordinates and rotational rate are made more precise during processing of observations using the least squares method for the 8-day interval.

Universal time is derived by comparing the results of the ongoing determination of orbits and their ephemerides, computed using data for the parameters of Earth's rotation, which were coordinated at the start with data from the International Earth Rotation Service. Thus, when each daily technical cycle of ephemeris information for each satellite is computed, the parameters of X_p , Y_p , and (UT1 – UTC) are evaluated. Averaging the data obtained for all of the satellites (with outlying values excluded) makes it possible to obtain more precise evaluation of the daily values for Earth's rotational parameters, which comprise the series of data measured at the ballistic center of the Glonass system. The daily Earth rotational parameter measurements are processed weekly. The results obtained are transmitted to the State Center of Measurement of the Earth's Rotation Parameters, where they are used to derive immediate and ultimate values.

Differential Methods of Navigation

The use of different variants of the differential method of navigation provides significant capability for obtaining highly accurate characteristics and improving the reliability of navigational measurements. The essence of the differential method involves using monitoring stations to identify and compute, in the form of corrections, strongly correlated components of the errors in navigational measurements. They determine the coordinates of their location on the basis of measurements of pseudorange and compare them with known results of their geodesic alignments. They use the results of the comparison to compute the appropriate corrections, which are transmitted along the communications channels to users of the Space Navigation System in the relevant region. Corrections allow users to make navigational measurements with enhanced precision.

There are four versions of the differential navigation method using the Space Navigation System's radio-navigation field: correction of position coordinates, correction of range, pseudosatellite system, and time correction.

In the location correction method, the monitoring station receiving equipment is used to determine corrections of the results of the navigational problem. The users themselves correct their navigational positions by the magnitude of these corrections. In this version, both the monitoring station and the users must operate with the same constellation of satellites, which requires that their work is coordinated.

The method of range correction involves the receiving equipment of the monitoring station to determine and transmit to the users corrections of the measured values of pseudorange for all visible navigation satellites. The users correct their measurement values of pseudorange by the appropriate correction factor obtained by the monitoring station.

The pseudosatellite system includes a ground-based transmitting device in the constellation of navigational satellites. The signals from this device are used for the user's navigational positioning. The receiving equipment of the differential station (in this version, the pseudosatellite) measures the correction factor for pseudorange of all of the satellites in the radio-visibility area and transmits them as part of the navigational message to the users. The user receives the signals from the pseudosatellite (differential station) along with the signals from the actual satellites and determines its coordinates considering the correction factor obtained from the pseudosatellite.

The time correction method involves having both the monitoring station and the user measure and track the time that signals arrive from the satellites. In this version, the user determines his location relative to the monitoring station using data on time and phase of signals received from the satellites at his own location and at the monitoring station, and also on geodesic coordinates of the station and the location of the satellites relative to it. It is assumed that if the coordinates of the monitoring station are precisely known, it is then possible to recompute the relative location of the user in geographic coordinates with enhanced precision.

It is hypothesized that the use of differential modes of navigation makes it possible to improve the accuracy of coordinate determination to 5–10 m and to increase the reliability of receiving navigational information by providing rapid transmission of signals containing information on incorrect functioning of the space system and the reliability of coverage of the working area by, in effect, increasing the number of satellites when the monitoring station functions as a pseudosatellite.

Use of the Glonass System

Space navigation became one of the first areas of applied cosmonautics that led to the extensive use of space navigation systems in the interests of society. Many current interesting projects implemented on Earth or in space would be unthinkable today without their use. Space navigation systems make it possible to develop and begin to use new navigation technologies for military purposes, for aviation and ships at sea; monitoring the orbits of satellites, launch vehicles, and upper stages and control of their motion; geodesy; geology; fishing; mining; land management; construction; transport; science; and other sectors.

Because positioning is achieved through uninterrogated measurements of pseudorange and pseudovelocity, the number of users of the Glonass space navigational system is unlimited and may be as large as possible. Users of the Glonass system are equipped with specially developed user equipment. At present in Russia, we are implementing a Federal program to develop various equipment prototypes for use by the Ministry of Transport and the Russian Aviation and Space Agency.

The use of the space navigation system by civilians makes it possible to improve safety and decrease the cost of operating air, sea, ground, and space vehicles by increasing the accuracy and reliability of navigational and time support. Plans call for using the space navigation system to solve all navigational problems, aside from those involving high precision aircraft landing and ship navigation on internal waterways. However, the use of the system in the differential mode enables solution of these problems, too, as well as many others, for example, those that arise in hydrographic, geodesic, and topographic work.

For air transport, global coverage of all airports in Glonass' operational zone makes it possible to improve local and general navigation, increase the density of air traffic, decrease the strain on major airports, and use backup airports more efficiently. Given the appropriate communications channels, Glonass can be used to prevent collisions and control air traffic. In particular, simply decreasing the length of each flight by several kilometers as a result of improved navigational support can save significant amounts of fuel for each aircraft.

At present there are proposals to use Glonass as an additional navigational system in air transport. It may be employed in instances where other systems could not be used. Here, it will be necessary to increase the reliability of navigational information receipt, that is, achieve virtually immediate detection and user notification of system malfunction, to ensure a reliable system work zone with the required geometric dilution of precision and low cost of user equipment. This type of use would increase the importance of employing the differential method.

The high accuracy of the system's navigational determinations makes it possible to solve all of the necessary problems of route navigation, seagoing navigation in the open sea and also sailing in narrows and in ports, as well as mining, fishing, and laying pipes and cable. The future use of Glonass in near-Earth space for navigational support of satellites and launch vehicles is promising. Russia has conducted successful tests of user equipment mounted on the Proton-M launcher equipped with a Briz-M upper stage.

The use of the Space Navigation System on land vehicles that have compatible communications receiving and transmitting equipment would make it possible to monitor traffic automatically along roads and exercise control over freight and passenger traffic when needed. The Space Navigation System can replace the majority of currently used radio-navigational systems for supporting industrial and scientific work in geodesy, geology, cartography, and other branches of science and technology. The system supplies highly accurate time information to stationary and moving objects on land, sea or in the air. We should also add less traditional uses of the Space Navigation System: for determining the orientation of objects when radio signals are received at distributed antennas, for synchronizing communications and power systems, and for monitoring deformations of Earth's surface in geodesy.

The main Russian developers of Navigational User Equipment are the Russian Scientific and Research Institute for Space Instrument Building, the Russian Institute of Radio-Navigation and Time, the Moscow Kompas Design Bureau, and others.

The use of the Glonass' radio-navigational field, in addition to solving major navigational problems, makes it possible to determine the location and velocity of the user in space, use the differential mode to perform high-accuracy local navigation relative to a ground-based correction station, perform high-precision geodesic alignment of ground-based objects that are far from each other, synchronize standards of frequency and time at remote ground-based facilities; and determine the orientation of an object based on radio-interferometric measurements received by distributed antennas.

Navigational System Integrity and Joint use of Glonass and GPS

The study of the requirements of various navigational system users suggests the need to identify and take account of integrity, an important characteristic of a navigational system. One measure of a navigational system's integrity is the probability of detecting a failure in less than threshold time. The Radiotechnical Commission on Aeronautics (RTCA) has introduced a definition of the integrity of a space navigational system that is either the user's primary means of navigation or an auxiliary means. If a navigational system is used as an auxiliary navigation aid, then integrity can be defined as the capability of the system to provide timely warning that it cannot be used for navigation in its current state. This means that the inadequacy of the system's function must be determined before errors in the output of navigational parameters exceed a specified threshold.

The integrity of a space navigation system that serves as the primary means of navigation requires that the system exclude inaccurate information from subsequent processing before errors in the output parameters exceed a stipulated threshold. Recently a great deal of emphasis has been placed on navigational system integrity, making this characteristic comparable in significance to the accuracy of systems and complexes. Situations associated with malfunctions of navigational systems can be either easy or difficult to detect. Easily detectable situations include signal loss or distortion and the occurrence in the navigational message of an indicator forbidding the use of the navigational signal. Situations that are difficult to detect include failures whose external manifestations are relatively indistinguishable from reliable performance. Failures of this type may have the following external manifestations: shift of the onboard timescale of the navigational signal, drift in the frequency of the onboard generator, drift of the carrier frequency of the output signal, and distortion in ephemeris information.

Various methods are used to monitor a space navigation system's integrity. Integrity monitoring occurs directly on board the navigation satellite or uses a ground-based monitoring segment. Monitoring stations receive signals from all visible system satellites and use them to determine whether one of the satellites has failed. Because it is not very efficient, this approach may not satisfy many users. Special equipment incorporating automated methods of system integrity

monitoring has been developed for such users. Such equipment is based on the use of redundant information obtained from other navigation aids, as well as redundant information from the space system itself. To implement these methods, various types of combined navigation devices have been developed, including altimeters that have a highly stable frequency standard and inertial navigational system combined with other radio-navigational systems. Incorporation of navigational equipment into a single functional, structural, and integrated system allows fuller use of the redundant information available onboard the moving object. To increase the level and degree of integration of navigation equipment as part of a unified navigational complex, the functions of various navigation and other radiotechnical systems have been combined, leading to the development of combined systems and multifunctional integrated complexes, as well as to integration of different technical devices measuring the same or functionally associated navigational parameters.

The development and use of combined receivers operating simultaneously with Glonass and GPS signals provides a significantly higher level of navigation service to air, sea and other users by integrating radio-navigational fields; increasing the number of radio signals received from all visible satellites, which may be 10–21; and improving the accuracy and integrity of the navigation system. To obtain the maximum effect from the integration of the Glonass and GPS systems, user receiving equipment at the level of radio signal processing system has undergone in-depth structural and functional integration.

The Glonass and GPS systems are similar, but from the standpoint of developing user equipment, they have a number of significant differences. For example, the two systems use close but not identical navigation signal frequencies. Thus, in the Glonass system, multichannel access to the navigation signals of each satellite is based on frequency separation; in GPS, this entails code-based distinction among signals. The systems also differ with respect to the navigational and auxiliary information that is transmitted in a navigational framework. Glonass systems time does not differ unintentionally from UTC(SU) by even a whole second. However, there is a constant difference of three hours, corresponding to the systems time of the ground control complexes between Glonass and UTC(SU). There is a few seconds difference between GPS systems time and UTC(SU) (13 seconds in 2001). This difference is measured during regular corrections of UTC(SU) times at the level of seconds. At the same time, there is no constant hour difference in the GPS system.

Both systems use similar, but not identical, systems of coordinates and time. GPS uses the world geodesic system, WGS-84, of the U.S. Defense Mapping Agency. This system has undergone several modifications to improve its accuracy, and at present, it virtually coincides with the Standard Terrestrial System (CTS) defined by the International Time Bureau (BIN) for the epoch of 1984. Glonass uses the geocentric system of coordinates, PZ-90. This system has its origin at the center of Earth. The OZ axis is aimed at the mean location of the North Pole for the epoch 1900–1905, as determined by the International Geodesic Union and International Geodesic Association. The OX axis is parallel to the equator for the epoch 1900–1905; the XOZ plane is parallel to the mean Greenwich meridian. The OY axis completes the coordinate system to the right.

The reference systems differ in that the WGS-84 system defines the 0Z axis as the axis passing through the current location of the pole, and the PZ-90 system defines the 0Z axis as passing through the mean location pole of the epoch from 1900–1905. This difference is recognized by including a transition matrix developed by the International Time Bureau, which may be used effectively to rectify the results of navigational measurements.

Another difference results from the fact that the system of coordinates, determined for each system by its network of observation points relative to which the “exact” ephemeris of each system is defined, has not been rectified, that is, there is no reliable information concerning the relationship between the two systems. Special studies would be required to establish the nature of this relationship.

Preliminary matrices for this rectification were obtained from experiments conducted by using the unpowered Etalon satellites launched as part of the program to ensure Glonass accuracy. The results showed that the following equations may be used to convert between the WGS-84 and PZ-90 coordinate systems:

$$\begin{bmatrix} X_{\text{wgs84}} \\ Y_{\text{wgs84}} \\ Z_{\text{wgs84}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1.4 \end{bmatrix} + \begin{bmatrix} 1 & -1.2 \times 10^{-6} & 0 \\ 1.2 \times 10^{-6} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & -\zeta \\ 0 & 1 & \eta \\ \zeta & -\eta & 1 \end{bmatrix} \times \begin{bmatrix} X_{\text{gl}} \\ Y_{\text{gl}} \\ Z_{\text{gl}} \end{bmatrix} \quad (2)$$

The equations were submitted to ICAO as preliminary proposals for developing a standard model of Earth for future systems of astronavigation and other uses. The discrepancy shows that the error resulting from the difference in the systems of coordinates is several meters. This must be considered in developing precision integrated user equipment that makes simultaneous use of Glonass and GPS signals. The equations cited are tentative and subject to further refinement.

System History and Developmental Phases

Modern space navigation was developed through the efforts of scientists and experts of many countries. The most decisive contributions were made by Soviet and U.S. specialists who developed the current low-orbital navigational systems for the USSR (Tsikada) and United States (Transit) and the mid-orbit USSR–Russian Glonass system and the U.S. GPS. The Soviet and U.S. systems were created through parallel development at virtually the same time, independently of each other. Each of these nations conducted development using their own scientific and technical facilities that had evolved in the process of previous development of radio-navigation technology, particularly during the period when first- and second-generation space navigation systems were being developed. The difference in the scientific traditions and approaches and in the level of development of technical principles led to differences in the designs and characteristics of the space navigation systems.

The first scientifically grounded proposal to use satellites for navigation was made in the USSR before the launch of the first satellite (10/4/1957).

Scientific research on the potential use of radio-astronomical methods for aircraft navigation, directed at defining technologies for creating aircraft radio-direction finders using intrinsic radiation from the Sun, Moon, and remote discrete sources was being conducted at the A.F. Mozhayskiy Leningrad Military Air Engineering Academy (LMAEA) between 1955 and 1957 under the direction of V.S. Shebshayevich. This research was limited by the small number of celestial sources of power sufficient to be received by antennas small enough to be installed onboard aircraft. In this context, the idea arose of using artificial celestial bodies equipped with radio transmitters of the required power. Rocket technology by this time had developed the potential for placing an artificial satellite in Earth orbit. In the research conducted, it was proposed to use such satellites as orbital radio-navigational points. It was emphasized that observation of orbiting points makes it possible to obtain a sufficient selection of variable navigational parameters (which would not be provided by the Sun and the Moon which move slowly in the sky). The conclusion of this work was that the use of satellites as carriers of radio-navigational signals was very promising. In October 1957, an interagency scientific and technical conference on the problem of radio-astronomical navigation supported proposals for the navigational use of satellites. They recommended comprehensive scientific research to develop rationales for various technical methods of developing satellite-based radio-navigation systems. The first publication defining ways to use satellites for solving aviation problems appeared in the USSR in the departmental (official bulletin) Information Collection No. 33, based on material presented at a scientific workshop held at the LMAEA in December 1957 (3).

The scientific and technical principles underlying low-orbital space navigation systems were developed significantly through comprehensive scientific research in a project entitled Satellite (1958–1959), which was performed by five Leningrad institutions (LMAEA, the Institute of Theoretical Astronomy of the USSR Academy of Sciences, the Institute of Electromechanics of the USSR Academy of Sciences, and two Naval Scientific Research Institutions) and the Gorkiy Scientific Research Institute for Radiophysics.

Over the course of 2 years, the scientific and technical foundations for the first generation Space Navigation System were laid. These included general structural principles for the system; methods of measuring (telemetry, radial-velocity, differential telemetry, goniometric-radio distance finding); selection of the frequency band to be used; the effect of conditions of radio-wave propagation; power for the radio links; a priori evaluation of the accuracy of positioning and flight altitude; the effect of instability of frequency and time standards; selection of satellite orbital parameters; evaluation of the accuracy of ephemeris prediction; a method for supplying users with ephemeris; techniques for coding the transmitted information; and specifics for using Space Navigation Systems for marine and aviation navigation.

Thus, from 1955–1959, Soviet researchers developed an independent scientific and technical foundation for satellite-based navigation and by the start of 1960, were prepared for practical development of a low orbital Space Navigational System, the implementation of which was turned over to the Leningrad Scientific Research Radiotechnical Institute (LSRRI)—currently the Russian Institute of Radio-Navigation and Time (RIRT). Here during the first half of the

1960s, the preliminary project was developed for the first Soviet low orbit Space Navigation System, Tsiklon.

Actual development work on the equipment for the first Soviet Space Navigation System was allocated as follows: the Research Industrial Association for Applied Mechanics in the city of Krasnoyarsk was the head organization and satellite developer under the direction of M.F. Reshetnev; the Moscow Research Industrial Association of Space Instrument Building was responsible for development of the onboard and ground-based radiotechnical complex and user equipment (directors, M.I. Borisenko and N.Ye. Ivanov); LSRRI was responsible for development of the user equipment and onboard frequency standards (directors, P.P. Dmitriyev and A.F. Smirnovskiy); and the Naval Scientific Research Institutes # 9 was responsible for development of applications for use of the system by the Navy. The Moscow Area Central Scientific Research Institute # 4 (subsequently the M.K. Tikhonravov Central Scientific Research Institute for Space Systems # 50) creatively interacted with all of the development organizations and worked on issues of control and ephemeris support of low-altitude Space Navigation Systems (V.A. Korobkin and A.V. Tsepelev).

The first navigational satellite, Kosmos-192, was inserted into orbit on 23 November, 1967. During this period, the developers faced numerous issues demanding theoretical interpretation and specific solution. Control of the satellite was performed in the Main Scientific Research Testing Center of Space Systems of the USSR Ministry of Defense (currently the Main Center for Testing and Controlling Space Systems of the RF Ministry of Defense). Starting in 1969, the Main Center was assigned the mission of increasing the accuracy of ephemeris support. By that time, it had become clear that the major source of error in navigational positioning was errors in the ephemeris. At that time, such errors were as high as 1–2 km. Multilevel work on this problem was directed by the head of the center's ballistic administration, V.D. Yastrebov. Under his direction, this problem was solved by E.V. Mesropov, V.M. Makarov, and G.M. Solovyev. V.I. Kudymov and B.I. Sukhikh of the Applied Mechanics Design Bureau also participated in this work. Success depended on many factors. First and foremost was that trajectory measurements by Doppler radio systems had to have stable accuracy, as well as highly accurate geodesic alignment. Further, an appropriate model of navigation satellite motion had to be developed to support accurate ephemeris prediction. Developers of radio-technical measurement systems from the Space Instrument Building Scientific Production Association were recruited to help in this work, as were geodesic specialists from Scientific Research Institute # 29 from the Military Topographic Administration and ballistic experts from Central Institute # 4 and the Krasnoyarsk Design Bureau for Applied Mechanics. As a result of their joint efforts, precision ephemeris prediction was attained that exceeded by a factor of 2–3 the precision stipulated in technical specifications for the Space Navigation System.

The first-generation space navigation system, Tsikada built and put into operation in 1979, consisted of four satellites inserted into circular orbits 1000 km high at an inclination of 83° and an even distribution of orbital planes along the equator. The use of the system enabled users, on the average of every hour and a half to two hours, to receive navigational signals in a session lasting for 5–6 minutes and determine their positions. The system used the Doppler

principle of positioning, in accordance with which during the session the Doppler shift of a highly stable satellite signal was determined and used to compute the user's coordinates.

Subsequently, Tsikada was equipped with instruments to detect vessels in distress that were equipped with radio-buoys emitting SOS signals at frequencies of 121 and 406 MHz. These signals were relayed to special ground stations where they computed the coordinates of the distressed ship. Equipping Tsikada satellites with equipment to detect such ships enabled development of the Kospas system. Jointly with the U.S.–French–Canadian Sarsat system, they established a united search and rescue system that saved several thousand lives.

Because of the discrete nature of the navigational sessions and their significant duration (5–10 minutes), the low-orbit space navigation systems could be used successfully as highly accurate navigation aids only for objects moving relatively slowly, especially ships at sea. In the late 1960, the problem arose of expanding the capabilities of the system to navigational support of faster moving objects and simultaneously increasing system efficiency and accuracy. It became necessary to create a universal navigation system, meeting the needs of all potential users: aviation, oceangoing vessels, ground transportation, spacecraft, ballistic and winged missiles, and space launch vehicles.

The Glonass system as a universal Space Navigation System for different types of users was developed during the 1970s. Experience gained in working to develop the low orbital system was fully used to develop high-accuracy intermediate-orbit space navigation systems. Scientific research accompanied development at all stages, starting with derivation of parameters during evaluation of technical proposals and ending with updating of software on the basis of test results. Research continued after the system went into use and is still continuing, now focusing on the further improvement of the system, increasing its accuracy and integrity, and expanding its functional capabilities.

The Glonass System was developed through a great deal of cooperation among organizations centered around the enterprises that had developed the first-generation Space Navigation System. The RF Ministry of Defense (the Military Space Force), the lead system customer, monitored its development and further improvement, and also deployment, support, and control of its orbital constellation. The M.F. Reshetnev Scientific Production Association of Applied Mechanics was the lead developer of the system, the navigational satellite, and the automated system for satellite control and its software. The Russian Scientific Research Institute for Space Instrument Building was the lead developer of the ground control system, the onboard equipment for navigation and command-measurement radio links and user equipment. The Russian Institute for Radio-Navigation and Time was the lead developer of the system synchronizing onboard and ground-based timekeeping devices and the user's navigational equipment. The Polet Production Association was the developer and manufacturer of the Glonass satellite. The Tikhonravov Scientific Research Institute for Space Systems # 50 and Scientific Research Institute for Military Topographic Management # 29 of the RF Ministry of Defense, RF Naval Scientific Research Institute # 9, and RF Air Force Scientific Research Institute # 30 all participated in developing the system, working on the statement of work, the scientific and technical supervision of testing, and validating the effectiveness

and principles for using the system. M.K. Tikhonravov Central Institute # 50 of the Military Space Force was assigned to improve and maintain the accuracy of ephemeris support of the system.

The Space Navigation System and the ground facilities of the Glonass system were developed by working groups of these enterprises led by Academician M.F. Reshetnev and Doctors of Technology Y.G. Guzhva, L.I. Gusev, A.G. Gevorkyan, N.Ye. Ivanov, A.V. Karpov, and A.V. Tsepelev. V.A. Korobkin was responsible for the development of the high accuracy model for computing ephemeris.

During the design and flight-test phase, the focus was on attaining the required accuracy characteristics for the navigation system. Specifically, developers were attempting to meet demands for levels of accuracy that had never yet been attained in any Russian space system. The low accuracy of ephemeris and frequency-time support that existed at the start of system design were the result of the following factors:

- a low level of geodesic and geodynamic support. The parameters of Earth's rotation were determined basically from data of the International Time Bureau, which failed to meet requirements for effectiveness and made the system dependent on the work schedule of the International Time Bureau.
- the level of stability of ground-based and small onboard frequency generators was more than an order of magnitude lower than that required.
- there were no measurement devices that could provide the required accuracy of trajectory measurements.
- the low precision of mathematical models of satellite motion for determining orbits and predicting ephemeris, resulting from uncertainties in defining the power of light pressure and computing active forces and force of the eccentric gravitational field.
- low precision of geodesic alignment of tracking devices.

The solution of the problem of attaining the requisite accuracy was based on implementing the following set of scientific and technical design and manufacturing measures:

- development of radio-technology measurement devices in the interrogating mode of the slant range with a potentially attainable accuracy of 1–3 m;
- development and deployment of quanta-optical systems for measuring slant range and angular coordinates with a maximum error of 1 m and 1 angular second, respectively;
- development of a radio-technology system for monitoring the phases of the navigational signal, linked with system frequency created by a highly stable ground-based standard generator;
- development of highly stable onboard generators of radio-signal frequencies;
- development of the Etalon passive satellite. Launching of passive satellites into orbit (Kosmos-1989 and Kosmos-2024) established that the forces acting on such satellites were approximately an order of magnitude lower than the

forces acting on actual satellites of the system. The use of passive satellites in the navigation satellite orbit made it possible to develop a reference point for increasing the level of geodesic and geodynamic support of the navigation satellites and identifying the nature of the effects of the unmodeled forces acting on them, as well as for conducting the first evaluation of the divergence between the system of coordinates used in Glonass and that used in GPS for computing ephemeris.

In addition to introducing new components into the system, a set of fundamental problems had to be solved, including developing and incorporating models of navigational satellite motion possessing minimal methodological error, although a significant number of disturbing factors remained unmodeled; development of methods and technical systems for determining (refining) orbital parameters and the parameters of the coordinated models of motion, including geodesic alignment of measurement devices, refinement of the parameters of Earth's gravitational field and the forces comprising light pressure; development of high accuracy methods for mathematical interpretation of measurements of current navigational parameters, techniques for standardizing highly accurate measurement systems and thus attaining the requisite accuracy characteristics; development of principles for monitoring the frequency–time characteristics and synchronization of onboard generators and methods for predicting their changes; development of techniques for geodesic and geodynamic support when using the ground control center's own devices and the potential of the developed software; and refining the coordinated parameters of the model of satellite motion and the geodesic alignment of the Command Tracking Stations and the quanto-optical stations based on the use of highly accurate measurements of the orbits of the navigation satellite and the Etalon passive satellite.

During the final development of the system, these problems were solved successfully. Theoretical principles and methods were developed that made possible efficient determination of the parameters of Earth's rotation from current satellite navigational parameters measured by the tracking stations. These methods provided more accurate determination of the kinematic parameters of satellite movement, with systematic errors of measurement by the Command Tracking Station and parameters of the model of light pressure forces and three parameters of Earth's rotation (change in duration of days and two component coordinates of Earth's pole). Comparison of the parameters thus obtained for Earth's rotation to analogous data published by the International Time Bureau showed that the discrepancy between them was less than one meter for the coordinates of the Earth's pole and less than one millisecond for the irregularity of Earth's rotation.

The data obtained for Earth's rotational parameters are used for extrapolating their values a month ahead. The error in extrapolation does not exceed two meters for the coordinates of the pole and 10 milliseconds for universal time. This level of accuracy satisfies current requirements for accuracy.

The next factor characterizing the precision of user positioning is the accuracy of synchronization of satellite navigational signals, which is performed using ground-based devices by aligning onboard timescales to the scale of a Central Synchronizer. Accuracy of synchronization is attained by creating and

using a highly stable ground-based generator—the major component of the Central Synchronizer with relative instability of 5×10^{-14} and the onboard standards of frequency with relative instability of $1 \times 10^{-13} - 3 \times 10^{-13}$ per day.

At the same time, the discrepancy between the phase of navigational signals must not have an error exceeding 10 nanoseconds, and prediction of changes in frequency-time corrections must be highly accurate. The use of generators with this level of instability is proposed in accordance with the program for developing Glonass systems.

Flight tests of the Glonass system were begun in October 1982 with the launch of the Kosmos-1413 satellite. By Order of the RF President on 24 September, 1993, the system, consisting of 12 navigational satellites, was adopted by the RF Armed Forces.

Figure 3 is a photograph of the group that participated in developing and testing the system after the final meeting of the State Commission in May 1991, at which the decision was made to compile documents for the President's signature to adopt the Glonass system for the Armed Forces. In the first row from left to right are Chairman of the Federal Commission, G.S. Titov, General System Designer, Deputy Chairman of the Federal Commission, M.F. Reshetnev, and Deputy General Designer, Yu.M. Knyazkin.

In 1995, the orbital constellation of the system was fully deployed. Further work on the space navigation system was conducted to increase the systems accuracy, ensure its integrity, and increase its service life. However, because of the sharp reduction in funding of new developments and expenditures on



Figure 3. The Glonass team.

replenishing the system, the size of the orbital groups was decreased. In 2001, the orbital constellation consisted of six functioning satellites. They were part of a Glonass/GPS joint use mode and methods were developed to use it in various areas. At present, resources have been allocated to supply the system and develop satellites with 5- and 7-year service lives. It is proposed that the system will be restored to its full complement by 2010. Measures have been mandated to increase accuracy by using onboard clocks with instability of 1×10^{-13} and a ground-based hydrogen standard of 1×10^{-14} , and also devices for comparing timescales of onboard clocks with the ground-based standard whose error is 3–5 nanoseconds. The potential of this system is far from exhausted. Its potential is in the continuous updating of the system and the development of in-depth integrated systems for controlling the motion of moving objects. It is considered that the Glonass system must function as an independent system and also in conjunction with GPS and other navigational systems being developed.

BIBLIOGRAPHY

1. Shebshayevich, V.S. *Radionavigatsia i Vremya*, 6–9 (1992).
2. Reshetnev, M.F. Information Bulletin of the Internavigation Science and Technology Center 1: 6–10 (1992).
3. Shebshayevich, V.S. (ed.). *Satellite Network Radio-Navigational Systems*. Radio i Svyaz, Moscow, 1993.
4. Volkov, N.M., N.Ye. Ivanov, V.A. Salishchev, and V.V. Tyubalin. *Uspekhi sovremennoy radioelektroniki*, 1, (1997).
5. *Glonass – Global Navigation Satellite System: Interface Control Document*. Moscow, 1998.
6. *Global Navigation Satellite System*. IPRZH, Moscow, 1998.
7. Solovyev, Yu.A., *Satellite-Based Navigation Systems*. EKO-TRENDZ, Moscow, 2000.
8. Kiselev, A.I., A.A. Medvedev, and V.A. Menshikov. *Cosmonautics on the Threshold of the Millennium. Conclusions and Prospects*. Mashinostroyeniye, Moscow, 2001.

VALERIY A. MENSHIKOV
G.M. SOLOVYEV
Khrunichev Space Center
Moscow, Russia

GLOBAL POSITIONING SYSTEM (GPS)

Introduction

The Global Positioning System, commonly referred to as GPS, is a worldwide, satellite-based positioning and timing system that allows suitably equipped radio receivers to locate themselves in four dimensions, latitude, longitude, altitude, and time, anywhere there is a reasonably clear view of the sky. The system is also

known as NAVSTAR, a convenient nickname that is not an acronym. The GPS system was developed, deployed, and is currently operated by the U.S. Air Force. GPS enables precision weapon delivery for all branches of the U.S. Department of Defense, as well as allied nations. Additionally, GPS supports civilian positioning and was always intended to support civil operations. The complete satellite constellation and ground support equipment that make up GPS was declared “operational” in December 1994, although civil use of the developmental signals started in the early 1980s. Initially, the civil signal was deliberately perturbed to prevent hostile use; this greatly degraded the civilian signal accuracy. This perturbation was called selective availability (SA), but the widespread advent of differential GPS, which calibrated these errors in real time, rendered this totally ineffective. In 1996, the President ordered this perturbation stopped, pending justification from the Department of Defense. In 2000, the selective availability perturbations in the signal were completely removed.

The fundamental operation is as follows: the 24 GPS satellites (see Fig. 1) are uploaded from the ground with their current and predicted positions (called ephemeris or orbital parameters). Small corrections of their space-borne atomic clocks are also uploaded. This information is broadcast to the user as a data modulation on an L-band signal (1575 MHz for most civilian users) that doubles

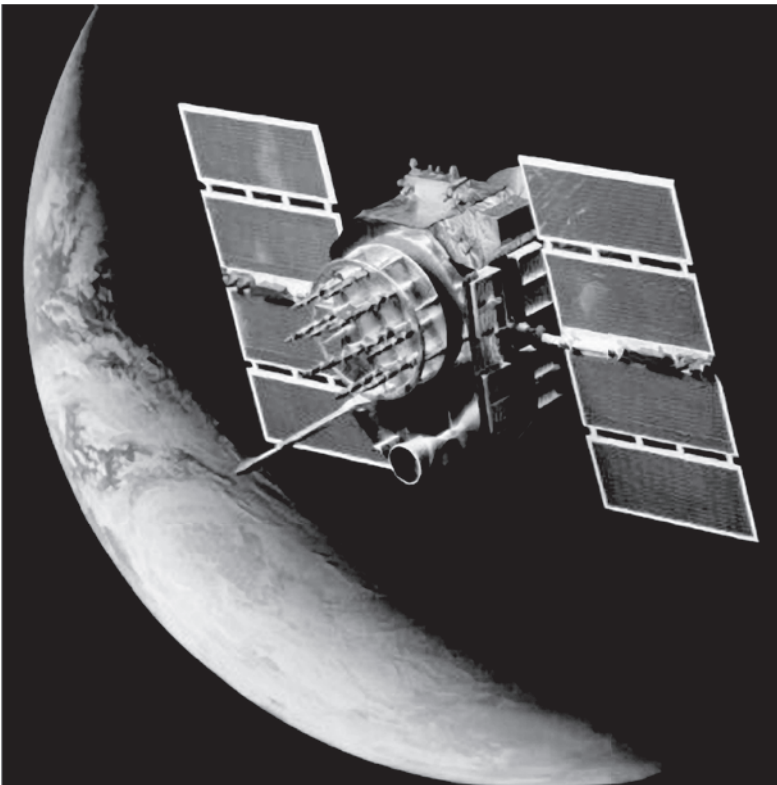


Figure 1. Early GPS satellite. A phase one GPS satellite built by Rockwell (now Boeing). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

as a precise, one-way ranging signal. Ranging is achieved by synchronizing the start time of a pseudorandom sequence of bits transmitted from the GPS satellites at an accuracy of about one nanosecond (10^{-9} s). Three very important results are achieved by this implementation. First, this makes GPS ranging a one-way signal that allows an infinite number of users to receive the signal and compute their position without saturating the GPS system. Additionally, this makes the GPS receiver passive, so that it does not radiate radio-frequency (RF) energy. Last, by receiving four or more satellite signals, users can synchronize their local clocks to GPS time, obviating the need for a very high quality and very expensive atomic clock in the receivers.

An important feature of the system is that all satellites broadcast on the same nominal frequency but use different modulation codes that are nearly orthogonal to each other. This technique is referred to as code division multiple access (CDMA). These codes are called pseudorandom noise or PRN codes. The user separates the signal from each satellite by correlating the incoming signal with an internally generated replica of the code for each of the satellites in view. The actual range measurement is the corrected difference between the phase of these codes and the local user's clock (called *pseudorange* because the true range is offset due to the local clock bias). Algebraically, four (or more) measurements allow the user to solve simultaneously for the four dimensions of location x , y , z , and t . More than four measurements allow improved accuracy and can also be used to monitor the integrity of the computed solution. A more detailed description of the Concept of Operations can be found in a later section.

GPS system accuracy, it can be shown, is a function of both the ranging accuracy from the satellites and the geometry of the satellite constellation being received. Typically 6 to 11 satellites are in view for users anywhere in the world who have clear views of the sky. Errors will be discussed later; typical civilian GPS positioning accuracies for nominal satellite geometries are summarized in Table 1. A full capacity GPS receiver can measure position, *velocity*, and *attitude* using multiple antennae. Thus, thirteen quantities can be measured by GPS: time (t) and the three dimensional position (x , y , z) and velocity (u , v , w), as well as the three attitude rotations (ψ , ϕ , θ) and the associated attitude rotational rates (p , q , r).

GPS is made up of three logically different systems that are commonly referred to as segments. The three fundamental GPS segments are the space segment, the ground control segment, and the user segment. The space segment (see Fig. 2) consists of approximately 24 satellites in six inclined orbital planes that have periods of 12 sidereal hours (11 hours, 56 minutes, and 4 seconds). Except for small perturbations, each satellite has a ground trace that is repeated twice per (sidereal) day. The corresponding altitude above the mean equatorial radius of the earth is 20,163 km. The orbits are nearly circular to keep the received power of the signal constant, and the orbital planes are nominally inclined 55° to the equator. All operational satellites have been launched from Kennedy Space Flight Center on Delta rockets, but the advent of the evolved expendable launch vehicle (EELV) will cause a switch to those vehicles. The space segment receives uploaded predictions of location and time corrections from the control segment and stores them for transmission to users. Three navigation signals are currently being broadcast: a civilian signal (called L1C) at 1575.42 MHz that has a modulation bit rate of 1.023 MHz; a military signal (called L1P/Y) also at

Table 1. Nominal GPS Median Accuracies for Civilian Users^a

Dimension	Operation					
	Nominal	Local differential	Wide-area differential	Carrier differential	Survey	Time transfer
Horizontal	10 m	0.5 m	1.0 m	0.01 m	0.001 m	NA
Vertical	20 m	1.0 m	2.0 m	0.02 m	0.002 m	NA
Time	50 ns	NA	NA	NA	NA	3 ns

^aThis table displays the GPS accuracies for civilian users with four or more satellites in view and reasonable geometries as a function of type of receiver and aiding. Nominal accuracies are indicative of a stand-alone, single-frequency code receiver. Differential aiding improves the accuracy of the position but does not affect the time, and the improvement in position is a strong function of the distance from the differential station. Carrier phase techniques provide enormous gains in positioning accuracy but require additional time and computation to solve for the unknown carrier cycles between the differential station and the user.

1575.42 MHz that has a modulation bit rate of 10.23 MHz; and a military signal (called L2P/Y) at 1227.6 MHz that has a modulation bit rate of 10.23 MHz. Details of the signal structure are given in Operational Concepts.

The control segment consists of five or more monitor stations, four ground antenna upload stations, and the Operational Control Center (OCS—Located at Schriever AFB outside Colorado Springs, Colorado). A backup control center is planned for Vandenberg AFB, California. Each monitor station measures the ranges to all satellites in view, smooths these measurements, and transmits these data to the OCS for further processing. The OCS predicts future satellite locations and satellite clock corrections. These data are then appropriately formatted and sent to the upload stations for relay to the satellites. The information is retained in satellite memory and sent to users as part of the data modulation scheme, at 50 bits per second. GPS is designed to retain its functionality, albeit at a degraded level, in the unlikely event that the ground stations cannot upload the data to the satellites. Modernization plans for the GPS constellation include satellites that can communicate directly with each other at higher data rates. This will provide greater capability in the event of loss of ground contact.

The user segment consists of the receivers, which lock on the signal, demodulate the data, calculate the corrected ranges, and transform this into position, velocity, and time. Differential transmitting stations (see Fig. 3) are considered part of the user segment, even though some of these may be other satellites (e.g., Wide Area Augmentation System—WAAS, Fig. 4) or transmission towers operated by the government (e.g., National Differential GPS—NDGPS, Fig. 5).

A Brief History of GPS

For 6000 years, humans have been developing ways to navigate to remote destinations. Driven mostly by the desire to transport goods by ship, early navigators remained within sight of land using a technique known as “piloting” that relied on navigators’ recognition of coastal features. The magnetic compass

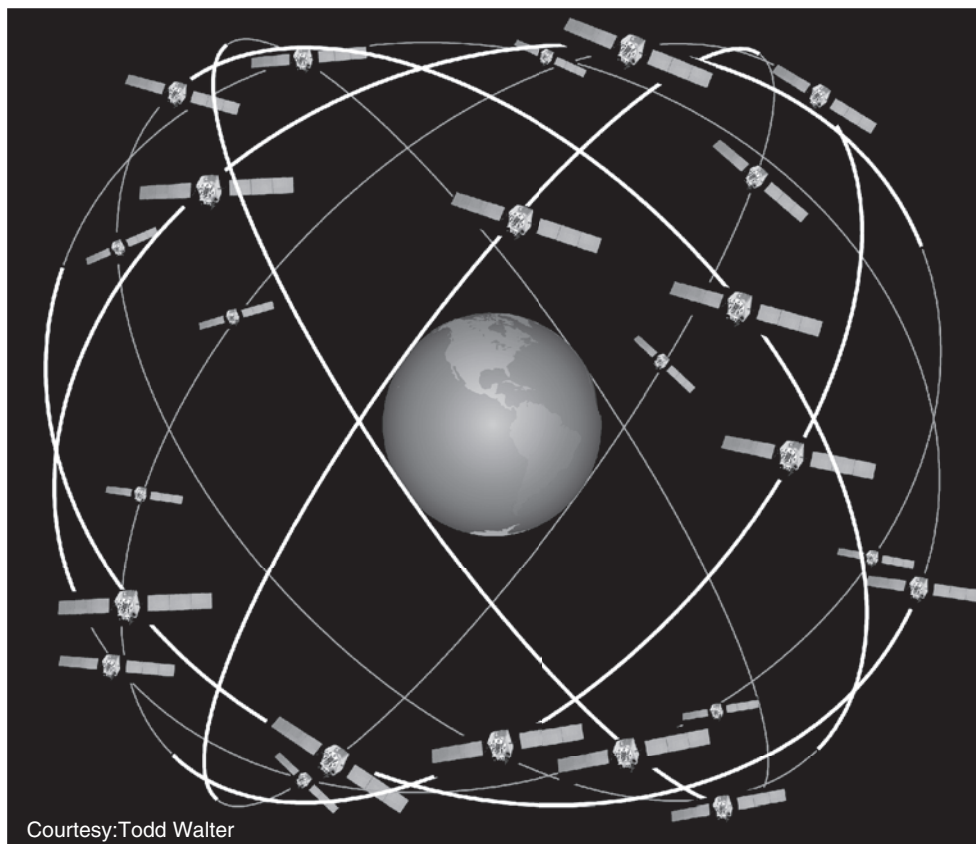


Figure 2. The GPS constellation consists of 24 satellites in six orbital planes. The orbital planes are nominally inclined at 55° and contain four satellites each. The satellites are not placed symmetrically around the orbital plane, but instead are placed in such a way that any single satellite failure has minimal impact on GPS. The orbital altitudes are 20,163 km above the equator. Some of the orbital planes may have extra satellites as on-orbit spares. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

appeared in China around the 1100 C.E. and in Europe approximately a century later. When forced to traverse a stretch of water outside the view of land or in inclement weather, navigators kept track of their position by “dead reckoning.” Navigators would record their heading and distance traveled by hourglass timing the passage of wooden logs thrown off the bow. Needless to say, the technique was notoriously inaccurate. The development of a sextant by 1731 (early versions existed in the thirteenth century) made determining latitude fairly routine. Early efforts to navigate precisely at sea led to so many deaths that in 1714 a King’s ransom was offered to anyone who could solve the problem of computing longitude (1). During the eighteenth and nineteenth centuries, the navies of the world refined optical instruments and timekeeping. This allowed reliance on the stars and planets to locate their ships precisely. These celestial navigation techniques fundamentally required angular measurements between

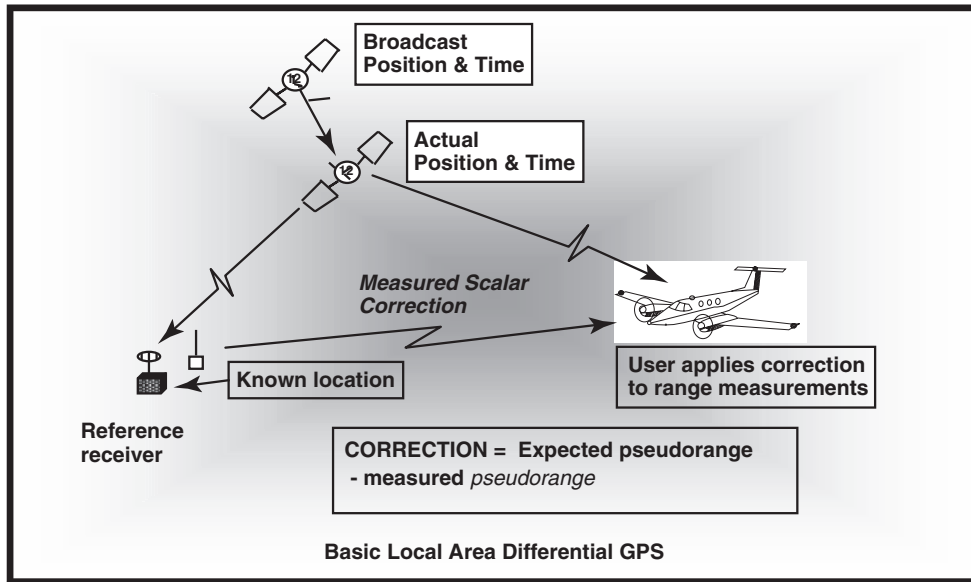


Figure 3. Block diagram of Basic Differential GPS. Differential corrections are broadcast to the user from a receiver in a known location that computes the correction from the difference between its known location and the GPS-measured position; hence the term “differential” corrections. The error at the reference receiver and the user are correlated across distance and time so that great improvement can be achieved across short distances and small time lags (typically 5–10 km and several minutes of latency). Note that the primary reason for using differential had been to reduce the effects of selective availability; when SA is off, much better accuracies and integrities are achieved. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the local horizon and the Sun, stars, or planets to find lines of position. Due to the motion of Earth, each angular measurement had to be carefully timed to attain the required accuracy. Earth’s “rim speed” at the equator is about 1500 km/h, or 24 km min; thus a 1-second error translates to about one-half mile. (On the other hand, because GPS uses the time of flight of a radio signal, a 1-second error for GPS translates into 300,000 km error in position.)

At the turn of the twentieth century, Marconi successfully transmitted radio waves across the Atlantic. By the 1930s, aircraft navigation was becoming a concern, and radio-navigation techniques were in their infancy. Early aircraft navigation aids consisted of direction-finding equipment, which gave a bearing to the transmitting station. Radio techniques such as radio beacons and LORAN were invented to overcome the limitations of celestial navigation and were largely deployed by the end of World War II. These techniques provided all-weather, 24-hour navigation service, but only within range of the signals.

GPS Predecessor: TRANSIT. In October 1957, the Soviet Union launched “Sputnik,” the world’s first space satellite. This triggered a flurry of activity within the United States to discover the exact details, especially the nature of its orbit. Two researchers, Dr. Guier and Dr. Wiefenbach, at the Johns Hopkins Applied Physics Laboratory (APL) had carefully studied Sputnik’s radio signal

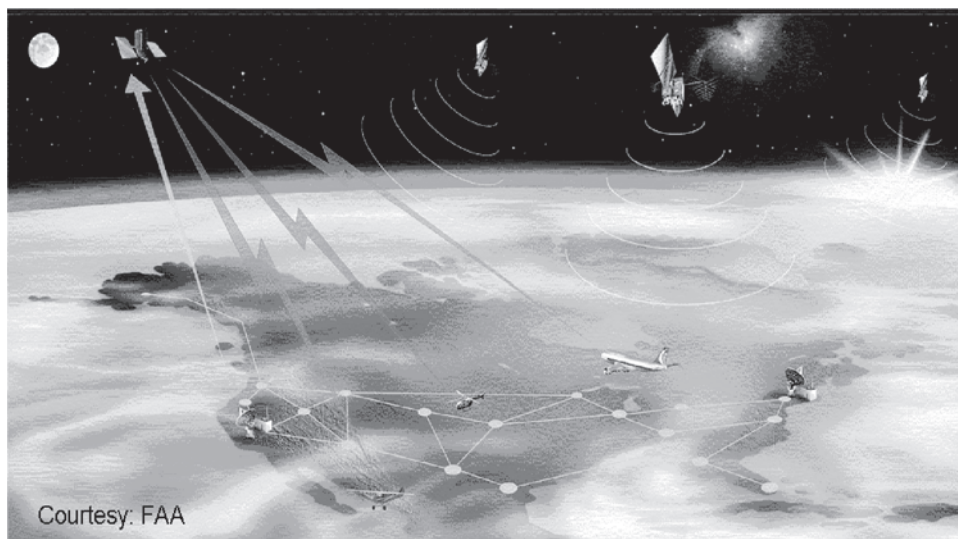


Figure 4. The Wide Area Augmentation System (WAAS) architecture. The basic functionality of the WAAS system is to use a widely spaced set of reference stations to produce a set of vector corrections for all users within the coverage space. Data is aggregated at each station and processed into a global set of corrections at redundant WAAS Master Stations (WMS) and in turn uplinked to satellites in GEO orbits. These satellites broadcast a message that allows users in the coverage area to compute their own corrections based on the WAAS data and rough knowledge of their own positions. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

and noted certain regular features. The most interesting of these features was the Doppler shift as the satellite passed overhead. This was caused by the change in the length of the line of sight and was enhanced by the satellite's high speed and low altitude. These scientists developed a computer program to determine Sputnik's orbit (2). Dr. McClure of APL, a colleague, realized that the problem could be turned on its head; the process could be reversed. By measuring the Doppler shift to a satellite of known orbit, listeners could calculate their own positions (3). This solved an important problem for the U.S. Navy that yielded precise all-weather positions for submarines and other ships. After speedy approval, a program was initiated under APL's management. The first two developmental TRANSIT satellites were launched by 1960, and the system became operational by 1964 (4).

TRANSIT eventually deployed an operational constellation that included about five polar orbiting satellites. They produced fixes every 35 to 100 minutes and provided horizontal accuracies of 100 meters or better for a stationary user. A moving receiver could compensate for velocity with some degradation in accuracy. TRANSIT was not generally used by aircraft due to the incompatibility of TRANSIT with the rapid platform motion of an aircraft. Additionally, aircraft require the third dimension (altitude) that the TRANSIT system did not provide. TRANSIT was, however, an important predecessor to GPS and pioneered a number of key technologies and concepts. TRANSIT led to a great refinement of

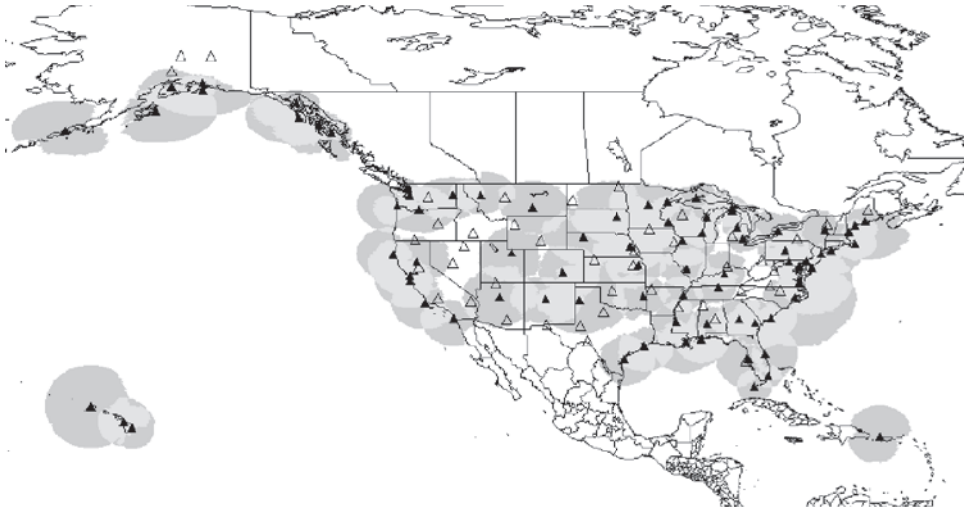


Figure 5. U.S. National Differential GPS System. This map shows the coverage of the NDGPS system as of 2001. Existing stations broadcast corrections in the 300 kHz band and generally have a range of 100–250 km. The current system covers the entire coastline and navigable rivers. Future upgrades are being deployed that will remove the gaps in coverage of the entire continental United States. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Earth's gravity field model, successfully tested dual-frequency correction techniques for ionospheric induced delays, and was crucial in developing stable and reliable frequency sources. TRANSIT provided only periodic updates and the degradation for a moving user made it unsuitable for aircraft. By the late 1960s, better systems were being explored by the Navy.

Additional GPS Predecessors: Timation and 621B. Timation was a program under the Naval Research Laboratory (NRL) whose goal was orbiting very accurate clocks. These clocks were to be used to transfer precise time among various laboratories around Earth. Under certain circumstances, users could also determine their positions by using the Timation signal. The approach was somewhat different from TRANSIT in that the radio signal allowed direct ranging by using a technique known as side tone ranging. Two satellites had been launched prior to the approval of GPS phase one in 1973. After that date, the Timation research effort was folded into the GPS development program. The NRL expertise played a key role in developing the atomic clocks used on GPS (5).

The third predecessor to GPS was a U.S. Air Force program called 621B. This effort was directed by an office in the Advanced Plans group at the Air Force's Space and Missile Systems Organization (SAMSO) in El Segundo, California. This concept was strongly supported and advocated by Dr. Getting of the Aerospace Corporation. This program evolved directly into GPS, although not before significant modifications were made to the original U.S. Air Force-only concept. By 1972, 621B had already demonstrated operation of a new type of satellite ranging signal based on pseudorandom noise (PRN). Successful aircraft

tests had demonstrated the PRN technique using ground-based “simulated” satellites located on the floor of the New Mexican desert.

The PRN modulation used for ranging was essentially a repeated digital sequence of fairly random bits (ones or zeros) that possessed certain useful properties. The sequence could be generated by using a shift register or for shorter sequences, could be stored in very little memory. Given the limited capabilities of computers then, this was a crucial feature. A navigation user could detect the “phase” or start of the signal sequence and use this for determining the range to the satellite. The PRN signal also has powerful noise rejection features and can be detected even when its power density is less than one-hundredth that of ambient radio noise. Furthermore, all satellites could broadcast on the same nominal frequency because properly selected PRN codes were nearly orthogonal.

When “tuned in” to a particular PRN sequence, all other PRN sequences appear to the user as simple noise. The PRN sequence can be tracked even in the presence of large amounts of noise, so other signals on the same frequency do not generally jam the signal of interest. The ability to reject noise also implied a powerful ability to reject most forms of jamming or unintentional interference. In addition, a communication channel could be included by inverting groups of the repeated sequences at a slow rate (50 bits per second is used in GPS). This communications channel allowed the user to receive the ephemeris, clock, and health information directly as part of the single navigation signal. The original Air Force concept visualized several constellations of satellites in highly elliptical orbits with 24-hour periods. This constellation design allowed deploying the satellites gradually (for example, to cover North America first) but complicated signal tracking due to the very high line-of-sight accelerations. Initially, the concept relied on continuous signal generation on the ground with continuous monitoring and compensation for ionospheric delays.

Program 621B was the immediate predecessor of the GPS effort, but the program came perilously close to cancellation several times a 10 year perspective on the history of GPS is in Ref. 6). In the early 1970s, Dr. David Packard, the Deputy Secretary of Defense, instituted important changes at the Department of Defense. One of these changes was to encourage joint programs that had multiple service participation. It turned out that GPS was the first “Joint Service” program. The first program director was Col. (Dr.) Bradford Parkinson, one of the authors of this article (one of the other authors is Dr. Jim Spilker who played a lead role in the design of the GPS signal structure). Dr. Parkinson was assigned to 621B in November 1972 and was directed to gain approval for the concept validation phase of the Defense Navigation Satellite System (DNSS), as the new DOD satellite navigation system was originally known. After many briefings of senior personnel in the Pentagon, a Defense Systems Acquisition Review Council (DSARC) meeting was held in August 1973, at which Dr. Parkinson presented a brief on the Air Force 621B program (7) approval was denied.

Meanwhile, Dr. Parkinson had presented the concept to the Director of Defense Research and Engineering, Dr. Malcomb Currie, who quickly appreciated the value of a three-dimensional, continuous, 10-meter positioning system. After the failure to gain approval, Dr. Currie invited Dr. Parkinson into his office and asked him to rethink the system and to ensure that it truly was a Joint Program, one that incorporated the best technology and concepts across DOD. He

wanted a synthesis, a new all-encompassing concept. Dr. Parkinson assembled about 10 of his key program members in the halls of the Pentagon during Labor Day weekend 1973. The result was a new system concept that was later named GPS, or NAVSTAR. By mid-December 1973, the senior DOD officials had been briefed and the reconvened DSARC gave approval. By June of 1974, the satellites, ground control system, and user equipment was on contract.

The first GPS satellite was launched in February 1978 and led to successful validation of the concept. The subsequent operational satellites incorporated certain additional nonnavigation payloads, which enhanced their value, but also undoubtedly delayed full operation. At the same time, the U.S. Air Force was not comfortable with having to shoulder the whole financial burden for the program and attempted to cancel GPS at least three times. In each case, civilian leadership (including the editor of this *Encyclopedia*) overruled the suggestion. GPS was declared operational in December 1995, although both civilian and military users had been using the available developmental system for more than 10 years.

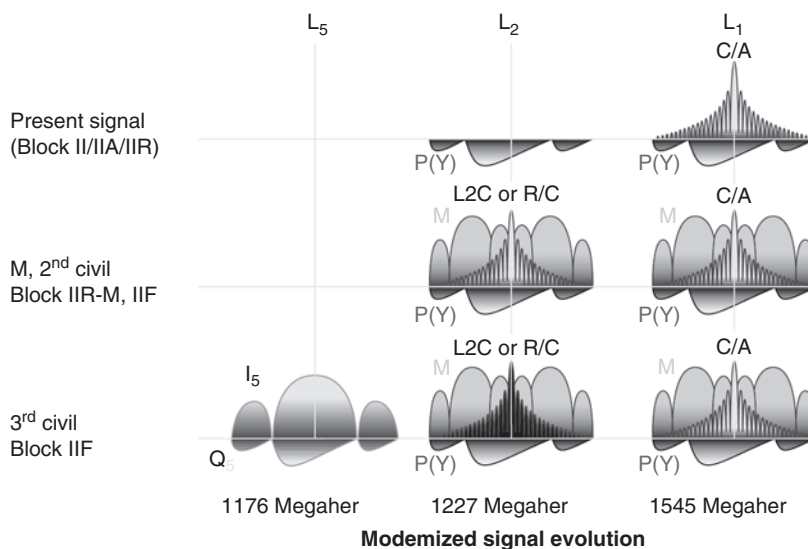
Due to the possibility that potential enemies might use GPS positioning against the United States or her allies, the civil signal was intentionally degraded through a process known as selective availability (SA). SA reduced accuracy for civilian users and remained part of GPS as a holdover from its original military history. SA was generally active, although ironically it was turned off during several national emergencies and international military campaigns due to the widespread military use of civilian receivers. It was slowly realized that the proliferation of differential corrections in the form of augmentations rendered these perturbations totally ineffective. As a result, a Presidential Decision Memorandum (PDM) was signed in 1996, which ordered the military to discontinue its use, pending justification from the DOD. In early 2000, SA was removed from the signals of all orbiting satellites.

During the first 25 years of GPS, several generations of satellite designs have been developed or are under development. These include I, II, IIA, IIR, IIRM, and IIF. In addition, there are plans for an upgraded version of GPS, known as GPS III, which is currently being defined. The Block IIRM and Block IIF satellites add additional civil GPS signals at other microwave band frequencies (see Fig. 6), which should materially improve the accuracy and robustness of the service (8).

GPS Concept of Operation

The design objectives of the GPS system were to provide a continuously available, worldwide, all-weather, three-dimensional precision navigation system for both military and civilian users on land, at sea, or in the air (or even in space). The GPS system had to operate, even on an accelerating platform such as a maneuvering aircraft or missile. Additionally, the system had to be passive, or one-way, so that it could service an unlimited number of users. As a military system, the signal is required to be both jam-resistant and antispoof.

Each of these requirements drives a certain set of constraints. To be worldwide and continuously available, only a satellite system can provide global



Courtesy: Aerospace Corporation

Figure 6. The GPS signal is undergoing modernization in preparation for GPS III. The current system is shown in the topmost frame with P/Y code on both the L1 and L2 frequencies, and the C/A (civil) code only on L1. The signal modernization calls for broadcasting a second copy of the C/A code on L2, and the military will get a new spread-spectrum code, called M-code, on both L1 and L2. The M-code is structured to broadcast most of its power into the nulls of the C/A code, maximizing spectral separation. A third civil frequency on L5 is set to be implemented on the late Block II-F satellites. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

coverage, especially over the oceans and polar regions. As a satellite system, frequencies less than 1 MHz skip off the ionosphere, and frequencies higher than 10 GHz are very heavily attenuated by atmospheric moisture. Satellite signal frequency was a compromise among accuracy (ionospheric delay), attenuation, and the power to be received by an omnidirectional user antenna. Thus, the selected signal was placed within the L band for best performance. Two additional constraints were established by the military: that the satellites could be totally serviced from the continental United States (CONUS) and that the constellation could be tested by using a small number of satellites to minimize project risk. These constraints led to satellites in MEO orbit, which costs significantly less in energy than a GEO orbit.

The quantitative requirements of the original GPS design were (1) to guide a bomb to within a 10 meter circle anywhere on the planet and (2) build an inexpensive (<\$10,000) device that could navigate.

Multilateration Positioning System. GPS functions as a multilateration, or rho-rho (ρ - ρ), system, that is, the range from at least three known locations is determined and the resulting intersection of the three spheres defines a single point that is the user location. In GPS, the system is complicated by the fact that the transmitters are moving and that the range cannot be measured directly. As

a simplification, assume that the GPS satellites are stationary and that the user is upon a flat nonrotating Earth. All of the satellites are synchronized and transmit a signal at the exact same time.

The user will receive the signal from each satellite at a different time due to the time of flight of the signal from the satellite to the user across the various ranges to each satellite. If the user possessed a very accurate clock that was time synchronized with the satellites, then the product of the time of flight and the speed of light would be the true range to the satellites. However, because the user is unlikely to have an atomic clock (a requirement which would make the receivers far too expensive), the user is not synchronized to GPS time. Thus, the measured range is offset by a consistent bias and is thus referred to as pseudo-range (ρ):

$$\rho_i = c \times t_i + b, \quad (1)$$

where c is the speed of light, t_i is the true arrival time, and b is the range equivalent bias in the user clock (time converted to meters). This measurement is taken simultaneously for each satellite. Even without knowing the exact time, the consistent solution for the ranges based on the user position and unique time bias can be computed. The general solution is nonlinear. The simplified equations for each satellite are

$$\rho_i = (x_u - x_i)^2 + (y_u - y_i)^2 + (z_u - z_i)^2 + b_u, \quad (2)$$

where the subscript i denotes each satellite and the subscript u denotes the users location and time (x, y, z are any convenient axes such as east, north, and up). Note that the satellite locations (x_i, y_i, z_i) are known from the navigation message on the signal. There are four unknowns in this equation (x_u, y_u, z_u , and b_u), and thus a minimum of four measurements is required to solve the equations. Generally, a direct solution is not computed, but rather the equations are linearized using a perturbation technique, and the position solution is computed using iterated least squares.

GPS Space Segment. The space segment of GPS is the satellite constellation (see Fig. 2) that consists of 24 or more vehicles in six orbital planes. The planes are inclined at 55° and are spaced 60° apart. There are four satellites in each of the orbital planes, but they are not evenly spaced. This was done to minimize the impact of any single satellite failure. Additionally, there are typically on-orbit spares in some of the six planes. The satellites are in a MEO orbit at a radius of 26,561.75 km (a mean equatorial altitude of 20,163 km). The orbits are almost perfectly circular and have an eccentricity of less than 0.01. The orbital period of these orbits is 12 hours of mean sidereal time (a mean sidereal day is the rotation of Earth to the same position with respect to inertial space, as opposed to a solar day, and is approximately four minutes shorter than a solar day). Thus, each GPS satellite repeats the same ground track, but passes the same location four minutes earlier each (solar) day.

The GPS payload consists of redundant atomic clocks, telemetry and control sections, and the signal generation subsystem. The atomic clocks are rubidium and/or cesium standards that typically have long-term stability of 1 part in 10^{13}

per day (or roughly a drift of 9 nanoseconds per day). The master control station monitors the atomic clock drift rates and models them as a quadratic,

$$\delta t = a_{f0} + a_{f1}(t - t_{0c}) + a_{f2}(t - t_{0c})^2 + \Delta t_r, \tag{3}$$

where t_{0c} refers to the master clock, t is the satellite clock, and the various parameters a_{f0} through a_{f2} are parameters for the polynomial fit to the satellite clock drift. The last term, Δt_r , compensates for relativistic effects caused by the motion of the satellites and their position within the gravity well, which has the effect of making the satellites gain 38 microseconds per day. This is compensated for by setting the main satellite frequency standard (10.23 MHz) slower by 0.00455 Hz. GPS is the first operational system known to require a correction for relativistic effects. All of these parameters are sent in the navigation message.

The satellites’ telemetry subsections are responsible for receiving the up-loaded navigation data from the Master Control Station (MCS). The data is encrypted before upload to ensure than no spoofing can occur. Internal status and health is also monitored and relayed back to the MCS. The signal generation subsection is detailed later in the discussion of signal structure. Currently, the nominal signal power is set at a minimum of -160 dBw for the Coarse Acquisition (C/A) code, -163 dBw for the L1 P/Y code, and -166 dBw for the L2 P/Y code, as shown in Table 2. Note that these power levels are well below the ambient noise level. From the satellites’ location, Earth subtends an angle of approximately 14° . A user at the limb of Earth is significantly farther away than one directly under the satellites. To compensate for this greater “space loss,” the antenna gain pattern on the GPS satellites is such that approximately 2.1 dB more gain is at the edges than at the boresight of the beam. The beam is also slightly wider than the 14° of Earth to allow non-GPS satellites on the other side of Earth to use GPS for positioning (9).

GPS Signal Structure. The PRN spread-spectrum coding that was originally pioneered for the Air Force 621B program contributes a great deal to the functionality of GPS. GPS uses a technique called *code division multiple access* (CDMA) such that each satellite broadcasts its message simultaneously on the same frequency, and yet the receiver can select each signal separately. The L1 signal is centered at 1575.42 MHz. This frequency is modulated with the satellite’s civilian PRN code using a biphase shift key (BPSK) modulation, that is, the phase of the carrier is reversed to indicate a “chip” transition (the military signal is in quadrature and the composite signal is called QPSK). A “chip” is the BPSK

Table 2. Minimum GPS Broadcast Power^a

Frequency	L1 (1575.42 MHz)	L2 (1227.60 MHz)
C/A	-160 dBw	N/A
P-code	-163 dBw	-166 dBw

^aThe specification for both the C/A code and the P/Y code (military) is such that the minimum broadcast power is well below the noise floor of the in-band radiation. Using the correlation properties of the PRN codes, a GPS receiver can reconstruct the phase of the signal and use this for position and temporal information.

analog of a bit, and in the L1C signal, is exactly 1540 carrier cycles long (or exactly $0.9775 \mu\text{s}$). The C/A (e.g., L1C) PRN codes are 1023 chips long, which means that the code repeats every millisecond. Last, the code itself is inverted every 20 ms to indicate a bit transition on the navigation message that is broadcast at 50 bits per second. An illustration of the signal structure is shown in Fig. 7.

The GPS C/A PRN codes are very carefully chosen for specific properties. The first property is that they can be easily generated by using a simple shift register. This was an important consideration during the development stages of GPS but is no longer relevant to modern CDMA design. The two main advantages of PRN codes are the signal spreading and the correlation properties. The unique properties of the set of PRN codes are that they have very good code-to-code and cross-correlation (multiple access) properties, even in the presence of large Doppler offsets.

If a PRN code is multiplied and integrated (i.e., correlated) against a local copy of itself, it produces a large correlation coefficient when the start (or phase) of the two codes line up. If the codes are out of phase, it produces a very small value. Likewise, correlating one of the PRN codes with a different PRN code produces a very small value for all relative phases. The implication of this is that

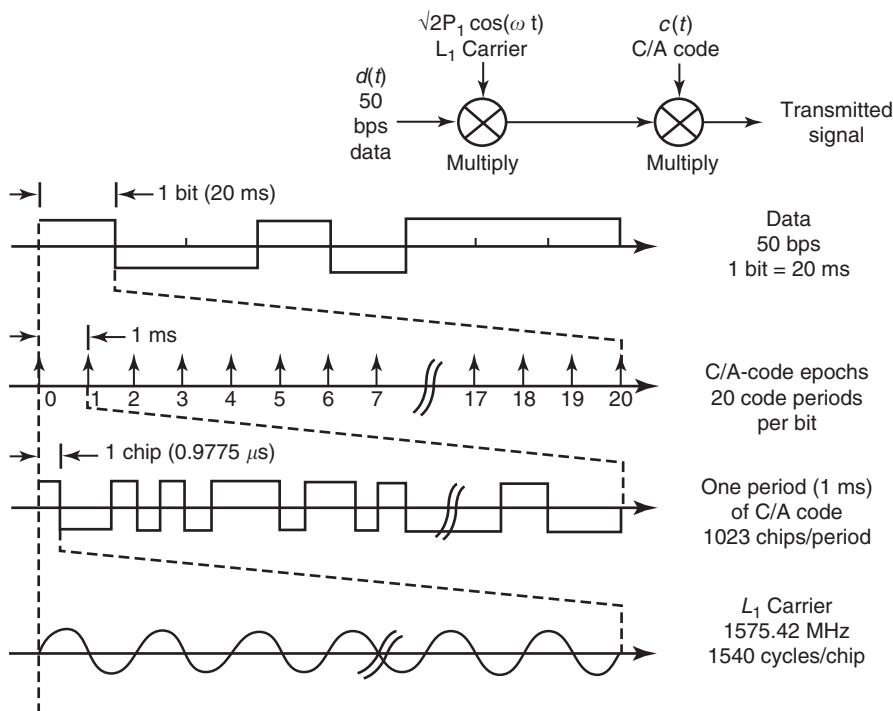


Figure 7. The GPS L1 signal structure is based on a carrier frequency at 1575.42 MHz. This frequency is modulated by biphase shift keying (BPSK) that uses phase reversal to indicate a changed “chip.” Each chip is 1540 cycles wide (or $0.9775 \mu\text{s}$). The code length is 1023 chips and repeats each ms. Last, every 20 ms, the entire code may invert to indicate a data bit reversal, modulated at 50 bits per second.

the phase of any given PRN signal within a receiver can be found by correlating a local copy at different phase offsets until a large signal is discovered. In this operation, all other PRN codes appear as noise. The worst case cross-correlation is -21.6 dB and is even lower at -23.8 if there is no Doppler offset. Pseudorange is the local clock reading (divided by the speed of light) at the start of the local code sequence when it is maximally correlated with the incoming signal (10).

Ground Control Segment. The GPS control segment consists of six or more monitoring stations around Earth, a Master Control Station (MCS), and upload ground-antenna stations. Each of the monitoring stations has a set of accurate atomic clocks and tracks both the code and carrier of each GPS satellite as it traverses overhead from horizon to horizon. The monitoring stations operate at both L1 and L2 frequencies to permit removing excess ionospheric delay. They also monitor atmospheric parameters such as temperature, atmospheric pressure, and humidity to permit estimating the tropospheric delay. By tracking the L-band carriers from horizon to horizon to a small fraction of a cycle (1% of an L2 carrier cycle is only 0.19 cm), a series of 15-minute averages is created and sent to the Master Control Station.

The Master Control Station receives the monitoring station tracking and ground antenna telemetry information and computes the current and predicted satellite clock offsets and satellite positions. It then converts this data to the navigation data formats described later. These rather complex satellite orbit/time filter estimating algorithms must also model the satellite solar radiation pressure, atmospheric drag on the satellite, Sun/Moon gravitational effects, including solid Earth and ocean tides, and Earth's geopotential model. Improved GPS satellite-to-satellite cross-link ranging data may also be used in the future. The navigation data are uploaded from several 10-m S-band ground antenna upload stations (11).

Navigation Data. The navigation data are encoded on the L1 C/A signal. This data message is transmitted at the rate of 50 bits per second and consists of a set of 6-second subframes (ten 30-bit words) and 30-second frames. The data encoded include the full ephemeris required to calculate the current satellite position, the satellite clock quadratic polynomial model and corrections to GPS time, almanac data used to position all the other satellites, and a hand-over-word for P/Y-code users. The almanac data allow a user to compute the rough positions of the satellite and thus narrow the search space both in terms of PRN codes and Doppler bins (12).

User Segment. The user segment or the GPS receiver is a very sophisticated digital signal tracking device that allows converting the faint signals from the GPS satellites into an accurate position solution. The GPS receiver must process the almanac (either stored or newly acquired) to generate a search space in terms of PRN codes and Doppler frequency bins. The incoming RF signal must be amplified, downconverted through an intermediate frequency (using a mixing process), and sampled into the digital domain. The PRN codes are correlated against the incoming digitized stream, and usually a delay lock loop (DLL) is implemented to keep the signal locked (13).

Once the signals are tracked, the corrections are applied to the raw pseudoranges, and the position and time bias are computed through an iterated least squares calculation. The positions are now reconverted to a useful coordinate

frame such as latitude, longitude, and altitude. The original GPS “manpack” receivers were backpack-sized devices that cost more than \$50,000 (see Fig. 8). GPS has benefited greatly from the semiconductor revolution, as has the typical consumer. A modern GPS receiver costs as little as \$100 and is small enough to be embedded into at least one wristwatch. Additionally, the computer that calculates the position solution can support many additional features such as map displays and waypoint guidance at minimal additional cost.

GPS Ranging Errors. There are several error sources that can corrupt the pseudorange and carrier phase measurements, as shown in Fig. 9. Thermal noise and interference effects degrade the performance of a typical receiver. Over the years, receivers have improved in noise performance. The free electrons in the ionosphere cause a code delay but a carrier advance (the so called code-carrier divergence). The ionosphere also varies in total electron count (TEC) depending on the state of solar activity and time of day. Delays are also associated with the troposphere that are a function of the slant range and moisture content below an



Figure 8. The original GPS “Manpack” cost more than \$50,000 each and was quite heavy. It did, however, satisfy the original mandate to produce an inexpensive device that could navigate. Modern-day receivers are much smaller and much less expensive. Today, one can buy both a watch and a cell-phone that has GPS built in. An inexpensive GPS receiver can be purchased for less than \$100. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

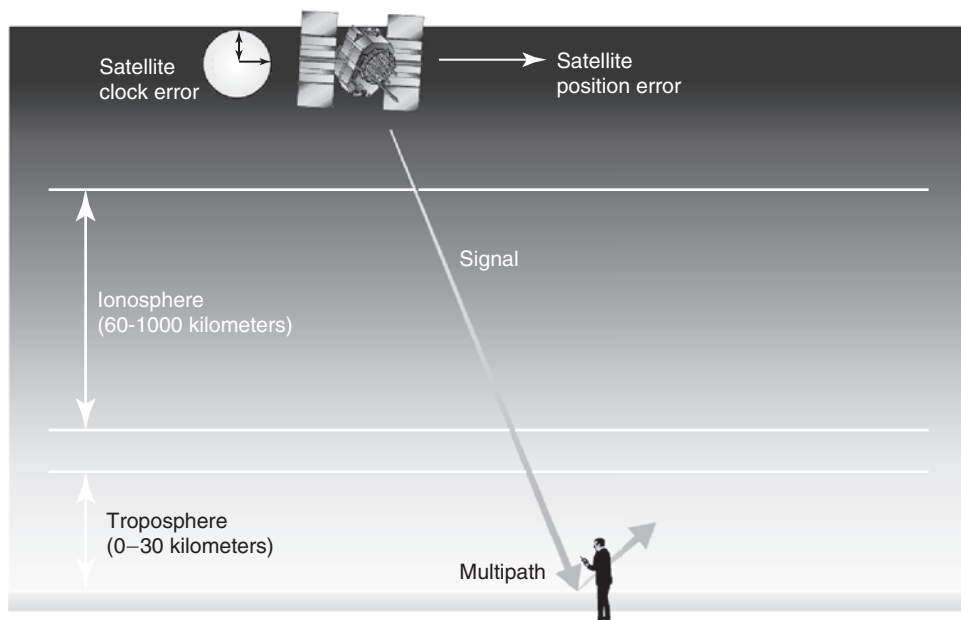


Figure 9. GPS ranging error sources. There are several different effects that can cause a ranging error in the GPS signal. Errors in either the satellite clock or orbital position (ephemeris) will cause errors. Additionally, both the ionosphere and troposphere cause delays in the signal. Last, multipath reflections of the signal can interfere with the original signal and distort the range information. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

altitude of 40 km. Errors in satellite position and clock directly cause errors in user ranging. Foliage can attenuate the signal, and more massive obstructions such as buildings or hills will block the signal completely. The latter is the origin of the urban canyon problem whereby GPS position is significantly degraded in cities that have tall buildings. User motion can cause the delay lock loops to be thrown off due to rapid changes in Doppler, though most terrestrial users will not experience such high rates of acceleration.

The largest error source (since selective availability has been turned off) is multipath. Multipath is the constructive or destructive interference with a reflected version of the signal that bounces off a nearby surface. Several techniques exist for multipath mitigation and are discussed in the next section.

GPS Error Analysis

To understand the potential of GPS, it is worthwhile to analyze the effect of the errors that occur when using it. In general, the errors are associated with measuring the range to the satellite. The ranges to four satellites must be processed to find the user's position, taking into account the locations of these satellites. Depending on geometry, the positioning error may be much

higher than the typical ranging error. The ratio of the *positioning* error to the *ranging* error is called the geometric dilution of precision (GDOP). If all ranging errors are zero mean, uncorrelated, and have the same variance, the general relationship is

$$\sigma_P = \sigma_R \bullet \text{DOP} \quad (4)$$

where

σ_P = positioning error

σ_R = ranging error

DOP = a multiplier due to geometry

The DOPs can be calculated by forming an array of unit vectors pointed at each satellite from the user's position, \bar{e}_j , using three convenient coordinate directions such as east, north, and up.

$$G = \begin{array}{c|c} \bar{e}_1^T & 1 \\ \hline \bar{e}_2^T & 1 \\ \hline \bar{e}_3^T & 1 \\ \hline \bar{e}_4^T & 1 \end{array} \quad (5)$$

The DOPs are then the square roots of the diagonal terms of the resulting 4×4 matrix:

$$\text{GDOP} = (G^T G)^{-1} \quad (6)$$

and

$$\text{Covariance(position)} = (G^T G)^{-1} \bullet \sigma_R^2. \quad (7)$$

The first three diagonal terms of GDOP refer to the coordinate directions selected above (e.g., east error factor, north error factor, and up error factor). The fourth diagonal term is the dilution for the range equivalent of the timing error. By dividing by the speed of light, one can change the value to the equivalent dilution in seconds (14).

The major sources of ranging error were discussed previously. Typical values are provided in Table 3. The typical dilution values (VDOP and HDOP) shown in Table 3 must be used with caution. If the satellite geometry is poor, it is not uncommon to find DOP multipliers of 10 or more. This is usually caused by a reduced number of satellites due to obstructions in the satellite line of sight. Typical causes are buildings, trees, and/or terrain. Modern receivers usually state the estimated error as part of the location message. Of course, the range of errors can be much greater than shown in Table 3, depending on age of update, atmospheric conditions, magnitude of multipath reflections, etc.

Table 3. Typical GPS Ranging Errors for Various Sources^a

Error source	Typical root-mean-square ranging errors single-frequency code-tracking user		
	High	Low	Typical
Ephemeris data	3.0	0.7	1
Satellite clock	0.5	3.0	0.9
Ionosphere (after modeling)	6.0	2.0	4
Troposphere	2.0	0.3	0.5
Multipath	15.0	0.2	1.2
Receiver measurement and noise	1.0	0.2	0.5
User equivalent range error (UERE)			4.4
Vertical rms error with VDOP of 3.0 = 13.2 meters			
Horizontal rms error with HDOP of 2.0 = 8.8 meters			

^aThe typical dilution values (VDOP and HDOP) shown above must be used with caution. If the satellite geometry is poor, it is not uncommon to find DOP multipliers of 10 or more. This is usually caused by obstructions in the satellite line of sight due to buildings, trees, or terrain. Modern receivers usually state the estimated error as part of the location message.

As can be seen in Table 3, the largest typical error is for ionosphere transmission delays even after modeling for a single frequency receiver. Ionospheric delays are caused by the interaction of free electrons in the ionosphere with the radio signal. One of the key observations is that most of the delay through the ionosphere is proportional to the inverse square of the carrier frequency. Thus, a dual-frequency user can directly estimate the ionospheric delay and substantially reduce or eliminate this error. Currently, only military receivers are truly dual frequency. These receivers have a current user equivalent ranging error (UERE) less than 2 meters, when multipath errors are small. Note that currently scheduled improvements in the GPS signal include two new civil signals at L2 (1227.6 MHz) and L5 (1176.45 MHz). By using the new second and third civil signals, all users will be able to calibrate the ionospheric delay directly. This is the largest error for most users, so accuracy will improve substantially as this error category is reduced to near zero (15).

For a number of years, the DOD deliberately perturbed the timing signal on GPS (a technique called selective availability or SA). This increased the UERE to about four times the typical values shown in Table 3. Of course, this also resulted in positioning errors that were about four times larger (16). The extensive use of real-time differential calibration of these errors made this technique ineffective, and it was discontinued by Presidential order. Additionally, over the years, the ground station has become much more skilled at calibrating the errors in the signal-in-space (i.e., ephemeris and satellite clock errors). Improvements in predicting orbits and clock drifts, plus increased uplink frequency, have reduced signal-in-space errors from 6 meters to less than 2 meters. This progression can be seen in Fig. 10.

The next largest category of error in Table 3 is multipath error. Multipath error is the misleading interference of the delayed reflection of the GPS signal. In

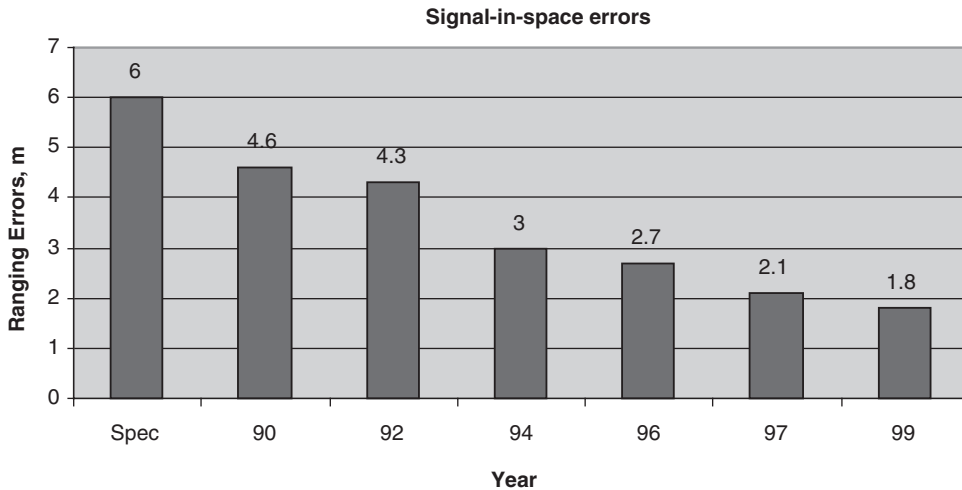


Figure 10. Signal-in-space errors have been steadily improving as ground segment operators have gained experience in orbit prediction and clock modeling. Additionally, the frequency of almanac updating has been increased. This has improved the signal-in-space error from the specification of 6 meters to less than 2 meters during a single decade. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

fact, this error will sometimes exceed the ionospheric error. Several techniques have been developed to mitigate the multipath problem. These range from better antenna designs whose gain patterns strongly attenuate signals coming from below the horizon to very narrow correlators that are immune to a large class of reflectors. Additionally, code measurements can be combined with carrier phase measurements that have a very different multipath response. As technology advances in receiver electronics and signal tracking, the receiver measurement noise will improve to the point of diminishing returns. Using dual-frequency receivers, multipath will be the dominant error source for GPS. Much of the future development in GPS receivers will be directed at eliminating the distortion from reflected signals (17).

Differential GPS

One technique used to augment GPS is known as “differential.” The basic idea is to locate one or more reference GPS receivers at known locations in users’ vicinities and calibrate ranging errors as they occur (see Fig. 3). These errors are transmitted to users in near real time. The errors (or their negative, which are corrections) are highly correlated across tens of kilometers and across many minutes. Use of such corrections can greatly improve accuracy and integrity. Several large-scale differential networks have been deployed in the United States and elsewhere (18).

Overview of DGPS Systems. The U.S. Coast Guard (USCG) within the United States and the International Association of Lighthouse Authorities

(IALA) have deployed a marine beacon differential system internationally, known in the United States as National Differential GPS (NDGPS). The Army Corps of Engineers is currently deploying additional beacons that are compatible with the U.S. Coast Guard differential system and cover the entire continental United States (see Fig. 5). The Federal Aviation Administration (FAA) is currently deploying the Wide Area Augmentation System (WAAS). WAAS is intended to provide enroute navigation and nonprecision approaches for aviation users (see Fig. 4).

The FAA is also developing a Local Area Augmentation System (LAAS) for Category I, II, and III precision landing capability at airports (see Fig. 11). This will require local ground monitoring stations to ensure the integrity of the system in addition to the nominal reference receivers. The U.S. Department of Defense is currently developing a new Military Landing System (MLS) to operate like the LAAS system but will be used on aircraft carriers and at forward bases. The system, called the Joint Precision Approach and Landing System (JPALS), has already demonstrated fully autonomous carrier landing using a specially equipped Navy F/A-18 (19).

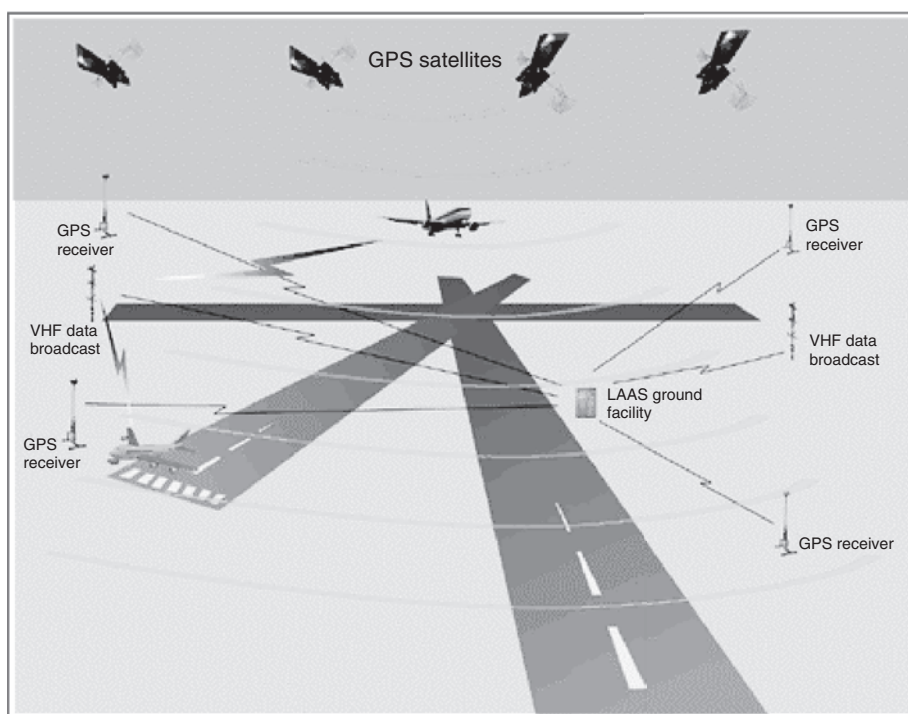


Figure 11. The LAAS system under current development by the FAA will provide precision approach capability using GPS. Due to the exacting requirements of Category II and III landings, the LAAS requires many cross-checks of the GPS system to ensure integrity. If one of these cross-checks fails, the time to alarm of the LAAS is specified at less than 6 seconds. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

There are many additional private and international systems under development or deployed. Various private companies sell their own proprietary carrier phase differential GPS systems for use in such diverse areas as construction, surveying, and archeology. Commercial wide area corrections are carried by at least one commercial C-band satellite broadcast, and several oil companies have put their own differential stations on oil drilling platforms to ensure accurate positions for the helicopters and ships that service these platforms. The next subsections will further explain these examples.

National Differential GPS (NDGPS). NDGPS is a system that has been developed primarily for marine use. Both the U.S. and European equivalent systems use marine radio beacons transmitting in the 300 KHz band as communication links for GPS corrections; ranges are of the order of 100 to 200 kilometers. The applications are mostly for ships operating in coastal waters or upon navigable rivers. Typical accuracies are of the order of 1 to 2 meters horizontally; many commercial GPS sets now offer small additional radios to receive these corrections. Although the initial deployment has focused on U.S. Coast Guard applications, both the Army Corps of Engineers and the Department of Transportation are extending the NDGPS system to cover the entire continental United States. To expedite this full rollout, the NDGPS system will take over Ground Wave Emergency Network (GWEN) transmission stations from the U.S. Air Force because these stations are no longer necessary and were to be decommissioned. Figure 5 shows the current nominal coverage of the NDGPS network (20).

Wide Area Augmentation System (WAAS). The WAAS, developed by the U.S. FAA, is specifically designed to ensure integrity and improve accuracy for civil aviation users. GPS, augmented by WAAS, offers its capability for both enroute navigation and nonprecision approaches (NPA). Fig. 4 shows the general architecture of the WAAS system.

GPS satellite data is received and processed at widely dispersed Wide-Area Reference Stations (WRS) that are strategically located to provide redundant coverage across the required WAAS area. Data is forwarded to redundant Wide-Area Master Stations (WMS) that process the data from multiple WRSs to determine the integrity, differential corrections, and residual errors for each monitored satellite and each predetermined ionospheric grid point. The multiple WMSs are provided to eliminate single point failures within the WAAS network. The differential corrections are allocated to satellite, clock, and ionosphere, so they are called “vector” corrections as distinguished from normal scalar corrections. Information from all WMSs is sent to GEO Uplink Subsystems (GUS), where it is uplinked to the GEO satellites. The GEO satellites downlink this data to the users via a GPS signal at the L1 frequency. Communication between ground-based stations (WRSs, WMSs, and GUSs) and other systems is accomplished via the Terrestrial Communications Subsystem (TCS), which provides two independent networks for redundant data communications among WAAS components.

WAAS accomplished the goal of very large area coverage by using a small number of widely spaced ground stations. No additional hardware is required on the user equipment because the data are modulated on an L1 signal. Additionally, the presence of a GPS-like ranging signal on GEO WAAS satellites can

improve the availability of the system if the WAAS signal-in-space has proper accuracy. Though the improvement in the positioning accuracy of WAAS is significant, more important is the bounding of worst case errors. Thus, the probability of hazardously misleading information (HMI) remains negligible ($<10^{-9}$). At the same time that WAAS bounds the worst case error, it does so without generating an unacceptable level of false alarms and keeping availability high at the nation's airports. As an added benefit, once the FAA declares WAAS operational, every single airport in the United States will potentially have NPA capability without installing any additional equipment at the airport (this will require new procedures).

WAAS has been under development since the mid-1990s and is currently in the final phase of deployment. The WAAS corrections, which are part of an additional GPS broadcast from two INMARSAT GEO satellites, are being extensively used. The typical accuracies for WAAS are shown in Table 4. The WAAS system is expected to be a boon to civil aviation in the United States, and both Europe and Japan are currently developing compatible nationwide augmentations for their own airspace (21,22).

Local Area Augmentation System (LAAS). Also being developed by the U.S. FAA is LAAS. It is designed to allow commercial aircraft landings down to Category II and possibly Category III. It is a highly redundant and reliable differential system that has several reference and monitoring stations and a very high standard of integrity. The LAAS system is meant to replace the current Instrument Landing System (ILS) at most large commercial airports (see Fig. 11). A Category III landing consists of a “zero-zero” landing (e.g., the visibility ceiling is at ground level and horizontal visibility is also zero). In practice, this means that the aircraft is landed by the autopilot. This is referred to as an “autocoupled” landing. Due to the automated nature of the landing, any landing system failure can be hazardous. This places an extremely high burden on LAAS to ensure that aircraft location is always within a well-defined error bound.

The LAAS system is designed with a very extensive set of cross-checks and verifications to ensure that no portion the system is operating outside its nominal parameters. These checks include validating the orbit of a given GPS satellite against a prediction based on the previous pass (from 12 hours before), checking the clock drift and both the range and range-rate of a rising satellite, plus many

Table 4. **WAAS Projected System Accuracies^a**

WAAS accuracies			
	50th percentile	95th percentile	99th percentile
Horizontal	1 meter	2 meters	5 meters
Vertical	2 meters	5 meters	10 meters

^aThe WAAS system was developed by the FAA to augment the GPS signal for civil aviation. The system is in its final stage of development, and many users are already using the corrections coming from the GEO satellites. WAAS excels in its ability to bound the worst case error and ensure that the probability of hazardously misleading information (HMI) remains very low while at the same time reducing the number of false alarms below the nuisance threshold.

other checks. Time to alarm is vital for protecting any landing aircraft from misleading information and is specified at less than 6 seconds (23,24).

Carrier Tracking Differential (CDGPS). Differential carrier tracking is another GPS technique that has been used by surveyors since the mid-1980s. By reconstructing the L-band radio-frequency (RF) carrier signal, a GPS receiver can attain tracking precisions of 1 to 10 millimeters. Specifically, a reference receiver (at a known location) measures the phase of the incoming carrier wave and transmits this information to a user. The user then compares this to the phase of the carrier wave received at the user's antenna. Because the wavelength of the L1 carrier is approximately 19 cm, a reasonable receiver can resolve this to 1% of the phase, or about 2 mm. Unfortunately, this is not accuracy. To attain equivalent accuracy, it is necessary to resolve the number of integer wavelengths along the RF path, that is, there are an unknown number of whole waves between the wave front arriving at the reference station and that at the user. Several techniques exist for resolving this integer cycle ambiguity. Satellite motion that can be exploited to do this differentially. This technique is referred to as real-time kinematic (RTK) GPS. When applied, this technique provides survey-level differential positioning whose accuracies are in millimeters. Thus one can locate an unknown point on the ground relative to a survey mark very rapidly and then maintain this accuracy as the user's receiver is moved. This is now being exploited for both construction survey and real-time, automatic, machine control (25).

The use of satellite motion can require some time to converge on the correct solution. An alternative for dual-frequency receivers is to set up a synthetic carrier wave by using the beat frequency of the L1 and L2 carriers together. The wavelength of the beat frequency is 86 cm, so the number of integer combinations to be searched in the position volume is typically much smaller and makes the problem more tractable. This technique is known as *wide laning*. Due to the advent of the two new civil frequencies on the block IIF satellites, users will be able to walk through a series of wide lanes to establish a carrier phase positioning solution very quickly (26).

Selected Applications

Applications of GPS have continued to multiply, as commercial and civil organizations apply creativity in using its capability. This section will not attempt to enumerate all current and future potential uses. Instead, selected examples will illustrate the revolutionary advances that have been made possible by this remarkable system. Many of the topics presented are at the cutting edge of current research and may yield profound improvement in our understanding of our world, as well as improved productivity and safety.

Survey and Crustal Motion. Until the advent of carrier phase differential GPS, measuring the relative distance or motion of large objects accurately over time required painstaking surveys using laser interferometry and tended to be one-dimensional. However, carrier phase differential GPS that can track 3-D relative positions down to millimeter levels across very long distances is revolutionizing the field of geomatics. Currently, experiments are underway that

monitor the relative positions of the mountainsides of several volcanoes in the states of Hawaii and Washington. Previous attempts at these kinds of experiments proved difficult due to the requirement for consistent line-of-sight measurements using optical sensors. Data recorded by using survey-quality GPS receivers have detected bulging of the mountains and are providing insights that may one day enable scientists to predict volcanic eruptions (27).

Similarly, hundreds of GPS receivers have been placed along fault lines throughout California and other parts of the world to validate theories about plate motion and gain valuable information on preconditions to earthquakes. Again, research in this area is still in its infancy, but it has never before been so economical or in some cases even possible, to measure the distance across large geographic features down to the millimeter level. At this time, data are being gathered to validate crustal motion models that will certainly lead to refinements in these models (28).

Aviation. The aviation industry has been an early adopter of GPS technologies and remains at the forefront of developing and implementing new GPS advances. In the early 1990s, a prototype GPS landing system for Category III (zero ft ceiling, zero miles visibility) was developed and demonstrated by Stanford University under an FAA grant. This system used carrier phase differential GPS to ensure a correct position. To resolve the integer cycle ambiguities quickly and robustly, two ground transmitters that broadcast GPS-like signals were used to augment the system. These “pseudolites” exhibited a large change in Doppler shift due to the rapid geometric change. The resulting system demonstrated more than 100 autocoupled landings at Crows Landing Airport in California; data were independently validated by using the Crows Landing laser tracker. The data showed an accuracy of better than 0.5 meter (3-D) in the final phase of landing (29).

During one of the autocoupled landings, a satellite upload from the Master Control Station caused the satellite to interrupt its transmission for approximately 1 millisecond. The Stanford system detected this glitch in the space segment and called off the landing in real time.

Though the FAA has not yet declared GPS operational as a *precision* navigation aid, most General Aviation and Commercial pilots use GPS as a backup system for navigation. Additionally, modern aviation GPS units are programmed with a full aviation database and can notify the user of airspace violations. In an emergency, these units can guide the pilot to the closest airport at the touch of a button.

GPS, as a full 13-state sensor for an aircraft, provides a powerful suite of information at a relatively low cost. Combined with inexpensive computer graphics, a synthetic “out-the-window” perspective display can be used to improve vastly the presentation of critical data to the pilot (30). The futuristic vision of tunnels-in-the-sky for improved navigation is being tested today in various laboratories around the world. Pilots who have experimented with these systems report a much reduced workload and greater situational awareness (31). The potential to reduce controlled flight into terrain (CFIT) could save many lives currently lost due to such accidents. Likewise, if all other aircraft are prominently displayed, it can reduce midair collisions. These displays have also shown great promise in enabling closely spaced parallel approaches (CSPA) in

inclement weather (32). This alone can save the United States billions of dollars in runway expansions and avoiding environmental impact that such construction would have on surrounding areas.

Vehicle Tracking. The so-called “urban canyon” can adversely affect GPS, but vehicle tracking remains a very important application. During urban canyon outages, most vehicle tracking systems use inertial augmentation to provide a position solution. Commercial companies have great interest in knowing where their equipment is currently located, and GPS provides an ideal answer. Many cities now have buses equipped with GPS receivers and radio transmitters. Each bus stop has a display of the current location of the next bus, and an estimate of the time to arrival. Likewise, many cities have GPS equipment on their emergency service vehicles to manage the response better. This has been shown very effective in reducing response time and managing these scarce resources during a large-scale disaster (33).

Vehicle tracking yields a great competitive advantage to a corporation. In one case, a cement company in Guadalajara, Mexico, would send fully loaded cement trucks into the city every morning, even though orders had not yet been placed. Using simple radio communication, this company responded to orders in less than half the time of any of its competitors. Though several trucks of cement would go to waste at the end of each day, within a short time, this company dominated the cement delivery market (34).

Last, law enforcement officials have been able to use GPS to increase their effective manpower by remotely monitoring suspects. After obtaining a court order allowing them to install a GPS receiver surreptitiously on a suspect’s car, Seattle police were able later to reconstruct the time and path of the location during a 2-week period, without alerting the suspect to the surveillance. This information led directly to evidence that convicted the suspect.

Precision Munitions. No discussion of GPS would be complete without a brief discussion of military applications. In spite of its explosive use for many civil applications, GPS was designed primarily as a military system, and to continue development, GPS must fulfill its primary mission. Several military applications for GPS were developed in recent years. An example is the JDAM. This precision-guided munition has demonstrated a battlefield accuracy of less than 10 meters. The trend in the future is to reduce the explosive warhead size of these kinds of munitions, which can be done only if the guidance system is capable of pinpoint accuracy (35).

On purely defensive military applications, the DOD recently deployed a Combat Survivor/Evader Locator (CSEL) radio for servicemen/women. This radio allows downed pilots to relay their positions to rescuers directly to enable rapid rescue and minimal exposure to hostile forces. The CSEL replaces four different individual devices with a single integrated package (36).

Space Applications. Some of the most innovative and unusual applications of GPS occur in the area of Earth sensing and space applications. Low Earth orbiting satellites can use GPS to measure both position and attitude. Precise satellite data can be used to refine gravitational models of Earth, and can be used as a sensor for attitude control. A soon-to-fly satellite experiment, the Gravity Probe B (GPB), uses very precise spherical gyroscopes to yield a quantitative measurement of Einstein’s theory of relativity. For the experiment to be valid,

GPB needs to fly a “drag-free” polar orbit to within 100 meters. GPS is used to provide guidance information to position the orbit of the satellite initially (37). Last, one of the most unusual applications of GPS is using the reflection of GPS signals from waves at sea to detect wave height in the open ocean (38).

Relationships to Galileo

Galileo is the European version of GPS. The European Union is committed to building a 30-satellite civil space-based navigation system at an estimated cost of 3.4 billion euros. The initial funding of 547 million euros is intended to fund the study and development phase, which is expected to take approximately 3 years. Galileo will be an entirely civil system that promises to be independent, but interoperable with the civil components of GPS.

Several outstanding issues must be resolved before Galileo becomes operational (planned for 2008). The most crucial is that the Galileo signals not interfere with any of the GPS signals. Ideally, Galileo would use a compatible geodetic reference frame and time base calibrated to GPS. This would present the Galileo satellites as an augmentation to the GPS constellation or conversely the GPS constellation as an augmentation to Galileo. Barring this level of interoperability, it is likely that Galileo will use a time base and geodetic reference frame distinct from GPS but one that can be easily translated back and forth between the two systems if real-time data are available. The exact configuration of the Galileo system is not yet certain and is the subject of current diplomatic negotiation between the United States and the European Union (39).

Future Improvements

The first block II-R GPS satellite was launched in 1997. Though the later versions of block IIs will be a bridge to a future GPS system, known as GPS III, the next generation of GPS is still being defined. Future improvements in the GPS system are driven by competing civil and military requirements. All users desire more signal power to ensure resistance to interference and/or jamming. In the last decade, GPS has become essential to virtually all DOD operations. International constraints on RF spectrum availability dictate that improvements remain within the radio navigation bands. On the civil side, the expectation has become that GPS will remain continuously available across the globe for the foreseeable future. Civilian users are urgently requesting the second and third frequencies to calibrate ionospheric delays and provide a backup if the L1 signal is jammed.

Several key advances are planned for the end of the block II series of satellites. The most important are two additional signals on the II-RMs and three on the II-Fs. The first additional signal is a replica of the C/A code but at the L2 frequency. This will allow direct measurement of ionospheric errors for civilian users. Military users will have a new split spectrum code (called M-code) on both L1 and L2. This code has the advantage of transmitting most of its power in the nulls of the C/A code, maximizing spectral separation. The signal modernization is shown in Fig. 6.

The II-Fs will include yet another civil signal at L5 (1176 MHz). This signal is intended to be a higher accuracy signal, which implies a higher chipping rate and a longer code sequence. Likely, it will include an unmodulated channel to enable much longer integration time for superior noise rejection. Other technical advances for the late II-Fs include intersatellite communication, as well as improvements in the rubidium/cesium clocks on board. Likewise, upgrades in the ground station facilities will reduce the errors in ephemeris predictions. For GPS III, the need for further increases in M-code power will probably lead to a spot beam of about 1000 kilometers (40).

Though all specifics of the GPS III concept are still to be determined, the United States intends to continue to provide and improve on a worldwide continuously available, precise, navigation signal that is free to all of the world. GPS III will undoubtedly continue in that tradition and provide a yet more robust and more accurate system of positioning on a global scale.

BIBLIOGRAPHY

1. Sobel, D. *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. Walker, New York, 1995.
2. Guier, W.H., and G.C. Weiffenbach. *John Hopkins APL Tech. Dig.* 18 (2): 178–181 (1997).
3. Piscane, V.L. *John Hopkins APL Tech. Dig.* 19 (1): 4–10 (1998).
4. Danchik, R.J. *John Hopkins APL Tech. Dig.* 19 (1): 18–26 (1998).
5. Parkinson, B.W. Introduction and heritage of NAVSTAR, the Global Positioning System. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.
6. Special Issue on Global Navigation Systems. *Proc. IEEE* 71 (10): (1983).
7. Parkinson, B.W., et al. Navigation. *J. Inst. Navigation* 42 (1): 109–164.
8. Swider, R. Department of Defense [online], 2000. Available: <http://www.igeb.gov/outreach/iberia-modernization.ppt>.
9. Misra, P., and P. Enge. *Global Position System: Signals, Measurements, and Performance*. Ganga-Jamuna Press, Lincoln, MA, 2001, pp. 284–287.
10. Spilker, J.J. GPS signal structure and theoretical performance. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.
11. Francisco, S.G. GPS operational control segment. In B.W. Parkinson, J.J. Spilker, P. Axelrad and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.
12. U.S. Air Force. Navstar GPS Space Segment/Navigation User Interfaces, [online]. ICD-GPS-200C, 1997. Available: http://gps.losangeles.af.mil/gpsarchives/1000-public/1300-LIB/documents/Other_Data/icdgps200c_irn1thru4.pdf.
13. Spilker, J.J. Fundamentals of signal tracking theory. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.
14. Parkinson, B.W. GPS error analysis. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.
15. Klobuchar, J.A. Ionospheric effects on GPS. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.

16. van Graas, F., and M.S. Braasch. Selective availability. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.
17. Braasch, M.S. Multipath effects. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. I. AIAA, Washington, DC, 1996.
18. Parkinson, B.W., and P. Enge. Differential GPS and integrity monitoring. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. II. AIAA, Washington, DC, 1996.
19. Jenkins, J. DCMilitary.com [online] 2001. Available: http://www.dcmilitary.com/navy/tester/7_30/national_news/18414-1.html.
20. U.S. Coast Guard. Nationwide DGPS Status Report. [online] 2001. Available: <http://www.navcen.uscg.gov/dgps/ndgps/default.htm>.
21. U.S. Federal Aviation Administration. WAAS, [online] 2002. Available: <http://gps.faa.gov/Programs/WAAS/waas.htm>.
22. Enge, P., and A.J. van Dierendonck. Wide Area Augmentation System. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. II. AIAA, Washington, DC, 1996.
23. U.S. Federal Aviation Administration. LAAS, [online] 2002. Available: <http://gps.faa.gov/Programs/LAAS/laas.htm>.
24. U.S. DOD and DOT. Federal RadioNavigation Plan, [online] 1999. Available: <http://avnwww.jccbi.gov/icasc/PDF/frp1999.pdf>.
25. Goad, C. Surveying with the Global Positioning System. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. II. AIAA, Washington, DC, 1996.
26. Hatch, R., J. Jung, P. Enge, and B. Pervan. *GPS Solutions* 3 (4): 1–9 (2000).
27. Roberts, C. *Measure & Map* 6: 28–31 (2000).
28. Bock, Y., S. Wdowinski, P. Fang, J. Zhang, S. Williams, H. Johnson, J. Behr, J. Genrich, J. Dean, M. van Domselaar, D. Agnew, F. Wyatt, K. Stark, B. Oral, K. Hudnut, R. King, T. Herring, S. Dinardo, W. Young, D. Jackson, and W. Gurtner. *J. Geophys. Res.* 102 (B8): 18013–18033 (1997).
29. Cohen, C., B. Pervan, H. Cobb, D. Lawrence, J. Powell, and B. Parkinson. Precision landing of aircraft using integrity beacons. In B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds), *Global Positioning System: Theory and Applications*, Vol. II. AIAA, Washington, DC, 1996.
30. Barrows, A., P. Enge, B. Parkinson, and J. Powell. *Proc. ION GPS* 1995, September 1995, Palm Springs, CA, pp. 1615–1622.
31. Jennings, C., K. Alter, A. Barrows, P. Enge, and J. Powell. *Proceedings of the ION GPS* 1999, September 1999, Nashville, TN, pp. 1923–1931.
32. Jennings, C., M. Charafeddine, J. Powell, S. Taamallah. *Proc. 21st Digital Avionics Syst. Conf.* 2002, Irvine, CA, 1B11/1–10.
33. U.S. Pat. 6374176 April 16, 2002, Schmier, K., and P. Freda, to NextBus Information Systems, Inc.
34. Katel, P. Wired Magazine [online] 1997. Available: <http://www.wired.com/wired/archive/5.07/cemex.html>.
35. U.S. Air Force, Factsheet [online] 2001. Available: <http://www.af.mil/news/factsheets/JDAM.html>.
36. U.S. DOD Joint Program Office [online] 2002. Available: <http://gps.losangeles.af.mil/csel/>.
37. Parkinson, B., and P. Axelrad. Closed loop orbit trim using GPS. *40th Int. Astronaut. Congr. Symp. Astrodynamics*, October 1989, Malaga, Spain.

38. Armatys, M., D. Masters, A. Komjathy, and P. Axelrad. Exploiting GPS as a new oceanographic remote sensing tool. *Proc. ION-NTM 2000*, Anaheim, CA.
39. European Union. Galileo Specification Document [online] 2001. Available: http://europa.eu.int/comm/space/doc_pdf/galileo_431.pdf.
40. Reaser, R. U.S. Air Force Joint Program Office [online] 2002. Available: http://www.ccit.edu.tw/~ccchang/Gps_modernization_ppt.pdf.

BRADFORD PARKINSON
JAMES SPILKER
GABRIEL ELKAIM
Stanford University
Stanford, California

H

HUBBLE SPACE TELESCOPE

Introduction

The Hubble Space Telescope (HST) is widely viewed as one of the most important scientific and technological achievements of modern times, comparable in its impact to Galileo's first use of the telescope for fundamental astronomical research in 1610. Although it is not the first astronomical observatory to exploit the benefits of viewing the Universe from outside Earth's atmosphere (Figure 1), it is the first to realize fully the gain in *clarity* of astronomical images that results from the absence of atmospheric turbulence. Without having to contend with the atmosphere's rapidly fluctuating refraction and transmission, the HST's angular resolution is limited primarily by light diffraction at the entrance aperture of its 2.4-meter telescope.

Earth's atmosphere glows from the emission of light by excited atoms and molecules. It is opaque at ultraviolet wavelengths below about 300 nm and strongly absorbs in broad intervals of the near-infrared band above 1100 nm. Outside the atmosphere, the optics of the Hubble telescope and its scientific instruments provide sharply focused and remarkably stable images against a very dark sky at wavelengths that span approximately 4.5 octaves—110 to 2500 nm. The ability to concentrate light from a point or compact source into a tightly focused image superposed on a dark, low-noise background allows the relatively small-aperture HST to detect extremely faint astronomical objects in its direct imaging mode—fainter by as much as 1.5 stellar magnitudes (four times fainter) than current 8–10 meter mountaintop telescopes.

This unique combination of capabilities has made HST one of the most productive scientific tools of modern times—and one of the most sought after. Observing time on Hubble is allocated by a process of competitive peer review on

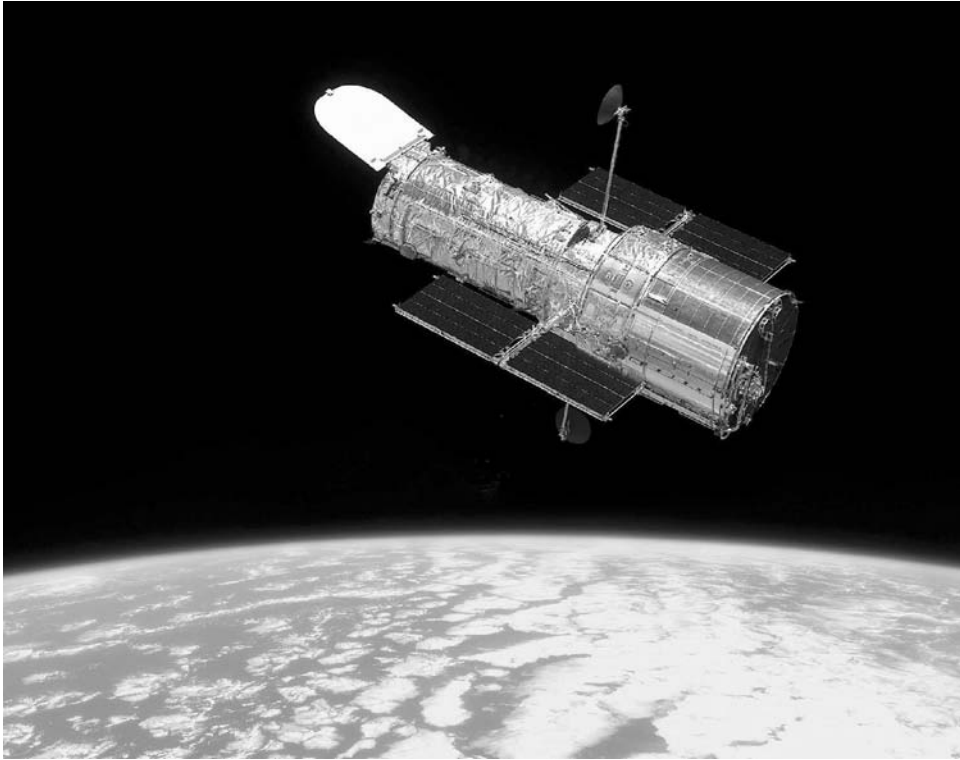


Figure 1. The Hubble Space Telescope, newly refurbished after Servicing Mission 3B in March 2002, orbits approximately 550 km above the surface of Earth. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the basis of scientific research proposals submitted yearly by astronomers from all over the world. The demand for using HST exceeds the available time typically by a factor of 6:1. The result is an almost continuous stream of amazing scientific accomplishments; many were unanticipated before Hubble's launch. These include the deepest view of the Universe ever acquired that revealed protogalaxies whose light was emitted when the Universe was less than 10% of its present age, the first demographic census of supermassive black holes at the centers of galaxies, accurate calibration of the age and expansion rate of the Universe, strong evidence acquired in partnership with ground-based observatories that cosmic expansion is accelerating, and frequent observation of dusty disks containing complex structures of rings and gaps possibly indicative of planet formation around other stars.

The HST is essentially unique among robotic space missions because it was designed for a long lifetime in space, enabled by regular orbital visits by crews aboard the Space Shuttle who implement technological upgrades of Hubble's instruments and other systems and perform a variety of maintenance and repair tasks on the spacecraft. This concept of preplanned, periodic servicing missions by Shuttle astronauts allowed the correction of a serious optical flaw in Hubble's telescope that was discovered shortly after it was first deployed in 1990. The first HST servicing mission in 1993 demonstrated that humans can carry out arduous

and complex work during a period of many days, encumbered by bulky spacesuits in the severe environment of low Earth orbit. Without the intervention of the Human Space Flight Program, the unmanned Hubble observatory would undoubtedly have come to be viewed by history as an embarrassing failure. Instead, Hubble became a national icon.

Hubble's History

The achievements of the Hubble Space Telescope are built on a legacy of scientific and technological progress that spans nearly a century (1). Its development and successful mission is the work of thousands of people. However, two individuals, Edwin P. Hubble (1889–1953) and Lyman Spitzer (1913–1997), stand out as seminal figures in the history that led to the HST. In the 1920s and 1930s, the observatory's namesake, Edwin P. Hubble (2), and his colleagues provided the first compelling observational evidence of two important properties of our Universe: the Universe is populated by a large number of widely separated individual galaxies of which our own Milky Way galaxy is but one; and the universe is systematically expanding, growing larger with time. These two discoveries set humankind on course toward our present understanding, as of the beginning of the twenty-first century, of the origin, composition, evolution, and fate of the universe and everything within it. These are the themes that dominate scientific research with the Hubble Space Telescope, and naming it in honor of Edwin Hubble is particularly apt.

Lyman Spitzer is universally recognized as the “father” of the Hubble Space Telescope. Although others had recognized the potential advantages of astronomical telescopes above the atmosphere, as early as the 1920s, Spitzer's classified study in 1946 for the RAND project (later to become the RAND Corporation) first articulated the scientific rationale that ultimately provided the underpinnings for modern space astronomy. Spitzer's greatest contribution, however, was that he championed the dream of a large telescope in space within the astronomical community, to the public, and to the federal government and relentlessly pursued that dream for the rest of his life (3). During the period 1965–1977, Spitzer and his collaborators, in particular astronomers John Bahcall and George Field, built the necessary scientific and political consensus that led the U.S. Congress in 1977 (fiscal year 1978) to initially authorize and fund the detailed design and development of what was then called “The Space Telescope (ST).”

In early design studies the ST had been conceived as a 3-meter telescope, commonly referred to as the Large Space Telescope (LST). In 1975, NASA reduced the proposed aperture to 2.4 meters to limit the program's expected costs. The reduced size of the telescope simplified its manufacture and testing, allowed the telescope and its surrounding spacecraft to fit within the envelope of the Space Shuttle's payload bay, and made it easier to achieve accurate and stable pointing of the orbiting telescope toward astronomical targets. The 2.4-meter aperture was judged sufficient to provide the light gathering capability and angular resolution necessary for the ST's most important scientific objectives, including detection of “standard candle” Cepheid variable stars in the galaxies of

the Virgo cluster, an essential step to providing a major improvement in our knowledge of the distance scale, rate of expansion, and age of the Universe (4).

The project's support in Congress rested in part on the understanding that tangible support would also come from outside the United States. Consequently, the Space Telescope was born as a collaboration between NASA and the European Space Agency (ESA). The agreement between the two agencies required ESA to provide the solar arrays from which the Space Telescope obtains its electrical power, a scientific instrument for the observatory (the faint object camera), and a portion of the staff required to operate the observatory after it was launched. In return, ESA was guaranteed a minimum of 15% of the observing time on the telescope for use by European astronomers.

NASA instituted a rather complex organizational structure, drawn from government, industry, and the academic world, to build and operate the Space Telescope. The Marshall Space Flight Center in Huntsville, Alabama, was designated as "lead center" for the development and launch phases of the mission and also was directly responsible for developing the optical telescope and the spacecraft. NASA selected the Lockheed Missiles and Space Company of Sunnyvale, California, to build the spacecraft and to integrate all of the flight hardware under contract to Marshall. The Perkin-Elmer Corporation of Danbury, Connecticut, was selected to design and fabricate the extremely precise telescope optics, fine guidance sensors and supporting structure, also under a Marshall contract. NASA assigned to the Goddard Space Flight Center in Greenbelt, Maryland, the responsibility for managing the development of the scientific instruments and the operational systems that would be used to command and control the Space Telescope from the ground. After launch and initial checkout of the integrated telescope, spacecraft, and scientific instruments, "lead center" responsibility for the long-term operations and servicing of the observatory was transferred to Goddard.

To design and develop the collection of sensitive instruments that would be the basic tools for astronomical research with the Space Telescope, in 1977, NASA selected five teams drawn from universities, industry, and government laboratories, as well as an interdisciplinary Science Working Group, to guide the overall project. After considerable deliberation, NASA concurred with the formally expressed desires of the astronomical community that the scientific program of the observatory be managed by an "independent" entity external to the government (5). Thus, in 1981, NASA signed a contract with the Association of Universities for Research in Astronomy (AURA) to create the Space Telescope Science Institute in Baltimore, Maryland.

The tumultuous early years of design and development work on the Space Telescope has been thoroughly (and grippingly) described by R.W. Smith (6). At the outset, the project was seriously underfunded and understaffed, and the technological problems to be surmounted were formidable. Although optical telescope technology applicable to the ST had reached an advanced state of development for national security applications, other required technologies were less mature circa 1980. Serious and costly problems were encountered, for example, in designing and fabricating the complex, lightweight, composite (graphite epoxy) structure on which the primary and secondary mirrors were mounted. Another challenging design problem centered on the extraordinary pointing

stability (± 0.007 seconds of arc rms) required for the telescope's line of sight for extended periods of time, to preserve the extremely high angular resolution provided by its tightly focused images (approximately 0.05 seconds of arc at visible wavelengths). When the ST program started in 1977, NASA anticipated an approximately 6-year development period, culminating in a launch on the Space Shuttle in the last quarter of 1983. This schedule proved far too aggressive, given the limitations of funding, the technical challenges, and the unwieldy management structure responsible for the observatory's development. Eventually the untenable nature of these problems was recognized. In 1983, NASA undertook a major reorganization of the ST program and provided a critical infusion of additional funding. At that time, the observatory was renamed Hubble Space Telescope (HST).

As the 1980s unfolded, the design, fabrication, and testing of all of the systems from which a complete HST observatory was to be assembled, never ceased to be severely challenging. Continuing technical problems, budget pressures, and schedule erosion compounded each other. One contributing factor described at the time was that the HST was developed as a "protoflight" unit. No budget existed for a "prototype" of the observatory to identify and resolve problems inherent in the design. Instead, these problems were first manifested in the actual flight hardware and that hardware frequently had to be redesigned and rebuilt as such problems arose. Nevertheless, the manufacture of the HST progressed. Testing of the scientific instruments was completed at the Goddard Space Flight Center in March 1984, and they were subsequently transported to Lockheed Missiles and Space Company (LMSC) in Sunnyvale, California. On 29 October 1984, the Optical Telescope Assembly (OTA) began its cross-country journey from the Perkin-Elmer facility in Danbury, Connecticut, to LMSC. It arrived at Moffett Field, California, on a "Super-Guppy" aircraft on 1 November. On 15 February 1985, the meticulous process of assembling the telescope, the spacecraft, and the scientific instruments into an integrated flight system was completed. The remainder of 1985 was devoted to an intensive and demanding test program of the assembled observatory.

NASA selected 21 June 1986 as the target date for delivering the completed HST to the Kennedy Space Center to begin preparations for its launch on Space Shuttle Atlantis on the following 18 August. On 6 January 1986, NASA announced a delay in the projected launch date of the HST to 27 October 1986. Three weeks later, on 28 January 1986, Space Shuttle Challenger and its crew were tragically lost, and the entire Shuttle fleet was grounded indefinitely. When it might be possible to begin the Hubble mission could no longer be foreseen and was no longer the primary concern of anyone associated with America's space program.

After years of anticipation and delay, expectations among astronomers, government officials, the press, and the public were unrestrained as Shuttle Discovery (STS-31) lifted off from Launch Pad 39B at the Kennedy Space Center at 8:33:51 A.M. on 24 April 1990, carrying the crown jewel of astronomy, the Hubble Space Telescope in its payload bay. On the following day, astronaut/astronomer Steve Hawley gently maneuvered the Shuttle's robotic arm (RMS) to lift the 12-ton observatory out of its moorings and raise it to a position high above Discovery. While the HST was still in the grasp of the RMS, ground controllers at the

Goddard Space Flight Center deployed the spacecraft's high-gain antennae and solar arrays and moved the arrays to the correct orientation to capture solar radiation for electrical power. After a brief checkout period, Hawley released HST from the remote arm. Several brief thruster firings separated Discovery from its payload, and at last the Hubble Space Telescope was orbiting freely, 380 miles above the surface of Earth. The aperture door was opened 24 hours later. For approximately 2 months thereafter, the process of verifying and calibrating the performance of the telescope, spacecraft, and scientific instruments continued.

The first test images of a field of stars, so-called "first light" images, were taken with two Hubble cameras, the wide field and planetary camera (WFPC) and the faint object camera (FOC) on 20 May 1990. As these simple images of individual stars were displayed on a monitor, it was immediately evident to the astronomers and engineers present that the telescope was badly out of focus. Several mechanisms were available to diagnose and correct optical problems within the telescope. A technique called "phase retrieval analysis" applied to the camera images, as well as measurements by the telescope's fine guidance and wavefront sensor interferometers, provided an accurate indication of the aberrations inherent in the optical image. To improve the image quality, the telescope's secondary mirror could be tilted, moved off-center, and moved in and out to adjust alignment and focus. In addition the 2.4-meter primary mirror was mounted on 24 pressure pads that could be moved individually for small adjustments to the mirror's shape. These capabilities had been included in the design to provide flexibility to remove almost any type of optical aberration that might be induced as the telescope experienced the transition from "1 g" to the weightlessness of orbit. The one type of aberration that could not be corrected in this manner was the simplest and least likely—spherical aberration. The entire project team was therefore stunned by its own conclusions that the Hubble telescope was afflicted with spherical aberration and that it could not be corrected using any of the onboard optical or mechanical systems. On 21 June 1990 NASA announced publicly that the Hubble Space Telescope was not working properly.

Subsequent investigation (7) revealed that the grinding process at Perkin-Elmer had removed too much glass, by a very small amount, from the primary mirror. Although the surface of the mirror was exquisitely smooth, it was too flat—the error reached approximately $2.2\mu\text{m}$ (about one-fiftieth the width of a human hair) at the outer radius. Consequently, light rays reflected from different concentric rings around the center of the primary mirror came to a focus at different locations (8). There was neither a single on-axis focal point, nor a single off-axis focal surface that defined the field of view. Instead, the focal points of different concentric rings were spread out across a range of 43 mm along the optical axis of the telescope. Rather than concentrating 70% of the light from a star in the central 0.1 arcsecond radius of its image as required, only 15% was contained in this tight central core. The remaining 85% of the light was widely dispersed in an "apron" around the central core of the image and was wasted. The error resulted from the improper assembly of a test device—a "reflective null corrector"—used to check the mirror's shape as the grinding process progressed. The error was evident in other, less precise test data, but these tests were discounted as unreliable themselves—in retrospect, a rationalization in an environment where the manufacturing process had fallen seriously behind schedule.

Despite intense criticism and negative publicity resulting from the announcement of Hubble's optical flaw, the HST program persevered in addressing the problem in several ways. First, the flaw itself was accurately characterized by both phase retrieval analysis of the astronomical images and by careful investigation of the reflective null corrector and other equipment with which the mirror had been manufactured at Perkin-Elmer. Even though the mirror had been ground to the wrong prescription, if one could accurately determine what that erroneous prescription was, it could be used to design highly effective corrective optics that might be incorporated in future HST scientific instruments. Several different diagnostic techniques gave the same, conclusive answer for the as-flown optical prescription (9).

Second, two instrument teams did incorporate this revised optical prescription in designs for instruments to be inserted into the telescope during the first servicing visit of Shuttle astronauts. Before HST's launch, a backup wide field and planetary camera (WFPC2) was already under development at NASA's Jet Propulsion Laboratory to ensure the observatory's long-term ability to acquire high-resolution wide-field images to meet its core scientific objectives. A relatively simple modification of small mirrors within WFPC2 would result in an accurate correction of the distorted telescope beam coming into the instrument. Concurrently, another team centered at the Space Telescope Science Institute and Ball Aerospace Corporation (10) invented an entirely new instrument called "COSTAR" (corrective optics for space telescope axial replacement). COSTAR contained pairs of small mirrors that could be inserted into the telescope beam to correct its spherical aberration before the beams passed through the entrance apertures of three other scientific instruments—the faint object camera (FOC), the Goddard high resolution spectrograph (GHRS), and the faint object spectrograph (FOS). These new optical designs would allow the telescope images to be sharply focused on the light-sensing detectors of WFPC2 and the other instruments. From the perspective of the scientific instruments, it would be as though spherical aberration had never occurred.

Third, although the image of a star or other "point source" was unfocused, it was very stable in the environment above Earth's atmosphere, and its shape could be accurately modeled. Even with spherical aberration, the image retained a very desirable property, a sharp central core that preserved some of the originally intended angular resolution. By applying a mathematical technique called "image deconvolution" (11), astronomers were able to remove the signature of spherical aberration partially from astronomical images, making it possible to implement a high-quality program of scientific observations during the early years of the mission before introducing corrective optics into the observatory (12–14). However, this early science was limited primarily to qualitative studies of relatively bright sources of minimal structural complexity. Hubble's highest priority objectives, including observing very faint objects far across the Universe and far back in time, observing individual stars in crowded fields, and performing quantitatively accurate measurements of an object's brightness, had to await the outcome of the first Hubble servicing mission.

On 2 December 1993, Space Shuttle Endeavor (STS-61) carried the first HST servicing crew into orbit (15–17). On board were the optically corrected WFPC2 and COSTAR, a new set of solar arrays redesigned to overcome

spacecraft jitter induced by the original arrays, new gyros, equipment to augment the spacecraft's computer, and several other components needed to improve Hubble's performance. The stakes of this mission were high. It was widely viewed as a demonstration that astronauts in space suits could do the kind of complex and taxing work necessary to build the future International Space Station. It was a demonstration that NASA could overcome past mistakes and failures. And for astronomers, it was the last opportunity to realize Lyman Spitzer's dream. News media and the general public around the world closely followed the mission by nearly continuous television coverage. The outcome is properly characterized as heroic. Each of the five scheduled days of extravehicular activity (EVA) was successfully completed, and the astronaut crew returned safely to Earth on 13 December. Two weeks later at the Space Telescope Science Institute, the WFPC2 science team and other astronomers nervously awaited the "first light" picture of a rich field of stars and broke into cheers when the tightly focused images were first displayed. At a press briefing on 12 January 1994, one of the designers of COSTAR, James Crocker, declared that the optical quality of the repaired HST was "as good as modern engineering permits and the laws of physics allow" (18,19). At that same briefing, Senator Barbara Mikulski of Maryland declared, "The trouble with Hubble is over!" In the years that followed, astronomers and the public as well were treated to the clearest and deepest views of the Universe ever experienced by humans—scenes of profound beauty and intellectual challenge.

Four fully successful servicing visits to Hubble have now been completed, encompassing 18 EVAs: SM2 (STS-82 on Shuttle Discovery) in February 1997 (20,21), SM3A (STS-103 on Shuttle Discovery) in December 1999 (22), and SM3B (STS-109 on Shuttle Columbia) in March 2002 (23). These have significantly modernized the technology of Hubble's scientific instruments and spacecraft systems. Each mission has left the HST at least an order of magnitude more capable scientifically and far more reliable functionally—as though a "new" observatory had been created. One more servicing mission is planned for Hubble, after which it will be at the apex of its scientific capabilities. The Space Shuttle will retrieve Hubble and return it to Earth at the end of its operational life in 2010.

Observatory Design

The HST observatory is comprised of orbiting hardware—the telescope, the spacecraft, and a set of scientific instruments (Figure 2)—and computers, software systems, and teams of people on the ground that monitor and control the orbiting equipment and execute the Hubble science program. The ground-based operations of the HST are described later in this article. The heart of the observatory is the orbiting optical telescope (24). Its 2.4-meter (7.9-foot) aperture provides an unobstructed light collecting area of $40,000 \text{ cm}^2$. The telescope is a classical "Cassegrain" design. Light collected by the primary mirror is reflected to a secondary mirror 0.3 m in diameter, mounted 4.9 m forward of the primary. The secondary mirror in turn reflects the converging light beam back toward the primary. The beam passes through a "donut hole" 0.6 m in diameter in the center

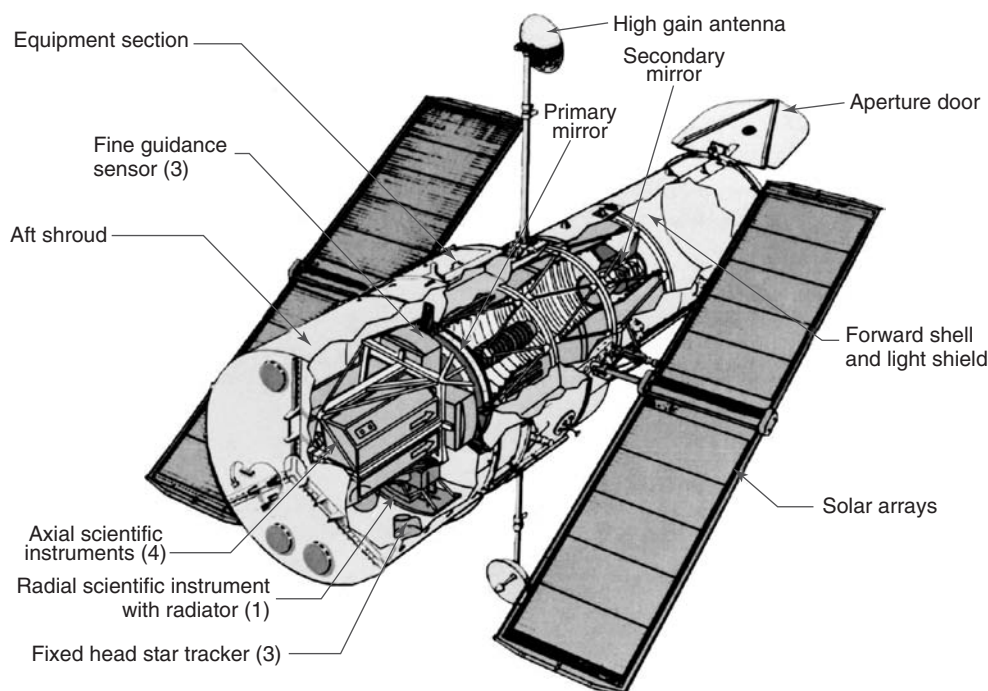


Figure 2. Cutaway schematic view of the major components of the Hubble Space Telescope. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

of the primary mirror and comes to a focus approximately 1.5 m behind the primary. This compact design folds the telescope of overall focal length 57.6 m into a package only 6.4 m long. The optical design of the telescope is also of a common type called “Ritchie–Chretien.” The surface of each mirror has the shape of a hyperboloid. A Ritchie–Chretien telescope corrects the focused image for both spherical aberration and coma (image elongation) across the entire field of view. Its focal surface is curved and subject to astigmatism, however, and these forms of optical aberration must be corrected within the optical systems of the scientific instruments that share the focal surface. A system of baffles, painted flat black, are mounted around the outside of the primary mirror, around its central “donut hole,” and around the secondary mirror, to attenuate stray or scattered light from bright, off-axis objects (e.g., the Sun, Moon, or bright Earth).

The telescope is encapsulated in the Hubble spacecraft (Support Systems Module). A light shield with an aperture door and a forward shell protect the telescope from the harsh thermal environment, from micrometeoroids and space debris, and from stray light (including sunlight). Behind the forward shell is the equipment section, consisting of an annulus of bays that contain approximately 90% of the spacecraft’s electronics, as well as the reaction wheels used to reorient the spacecraft from one pointing direction to another. At the back end of the spacecraft is the aft shroud that houses the focal plane assembly, the part of the optical telescope assembly’s graphite-epoxy structure in which the scientific instruments are mounted. All of the spacecraft’s interlocking shells, light shield,

forward shell, equipment section, and aft shroud, provide a benign thermal and physical environment, cloaked in darkness, in which sensitive telescope optics and scientific instruments can operate properly for many years. The spacecraft is about 13.3 m (43.5 ft) long, excluding the open aperture door, and its widest diameter is 4.3 m (14 feet), excluding the solar arrays. Combined, the spacecraft, telescope assembly, instruments, and other equipment weigh about 11,100 kg (24,500 lb).

During its mission, the HST has had three different sets of solar arrays—two large, rectangular “wings” containing solar cells that convert sunlight to the electrical power needed to operate all of the orbiting hardware systems and to keep Hubble’s six NiH_2 batteries charged for continuing operation during orbital night. The first two sets of arrays, provided by the European Space Agency, consisted of 48,760 silicon cells mounted on flexible blankets that could be unrolled somewhat like a window shade. The flexible design had the unintended effect of imparting small motions or “jitter” to the spacecraft resulting from sudden thermal flexure as it moved from the cold of orbital night to the heat of orbital day. Modifications of Hubble’s pointing control software mitigated this problem for the first set of arrays, and an improved design of the second set of ESA arrays, mounted on the HST during the first servicing mission in 1993, led to even better performance. During Servicing Mission 3B in 2002, an entirely new solar array design was introduced into the HST. These are smaller, mechanically rigid array wings comprised of gallium arsenide (GaAs) solar cells that are approximately 30% more efficient in converting sunlight to electricity than the prior arrays. When new, these arrays provide approximately 5700 Watts of electrical power. They should amply meet Hubble’s power needs to the end of its mission in 2010.

The exterior surface of Hubble experiences variations in temperature from -150 to $+200^\circ\text{F}$ in going from orbital night (Earth’s shadow) to orbital day. Its orbit, inclined at an angle of 28.5° to the Earth’s equator, precesses in space in a period of 55 days. So, the spacecraft goes through hot and cold seasons as Earth moves around the Sun and as the HST orbit precesses around Earth. Despite the harsh thermal environment, the interior of Hubble is maintained within a narrow range of temperature, in many areas at a “comfortable room temperature,” by its thermal control system. Temperature sensors, electric heaters, insulation inside the spacecraft and on its outer surface, and paints that have special thermal properties all work in concert to keep the equipment inside the spacecraft at proper operating temperatures.

To avoid blurring Hubble’s crystal clear images, the optical axis of the telescope must be tightly locked onto the selected astronomical target. Line-of-sight jitter is restrained by the spacecraft’s pointing control system to approximately ± 0.005 arcseconds rms during periods of 24 hours or longer. The full $\pm 3\sigma$ band of small motions, 0.030 arc seconds, corresponds to the angle subtended by a dime at a distance of 725 miles (1167 km), roughly the distance from Washington, D.C., to Chicago. Hubble must also be able to move from one target to another and place the new target within the aperture of a scientific instrument at an accuracy of 0.01 arcseconds. The exquisite pointing accuracy and stability are achieved by using a complex system of onboard sensors and actuators working together under the control of Hubble’s central computer and pointing control

system software. There is no propulsion system on the spacecraft. After completing observations at one pointing in the sky, the spacecraft is commanded to rotate in pitch, yaw, and roll, driven by adjustments to the spin rate (angular momentum) of motor-driven reaction wheels. There are four of these flywheels on board, rotating as fast as 3000 rpm; however, only three are required to move Hubble. During the slew to a new target, the change in orientation is measured relative to the orientation of three gyroscopes. These delicate and highly precise gyros also assist in stabilizing Hubble's pointing while it is acquiring and observing a celestial target. The spacecraft carries a total of six gyros to provide full redundancy because they have a limited lifetime. At the preprogrammed end of a three-axis slew the rotation of the reaction wheels is changed to brake the spacecraft's motion. Long electromagnets attached to Hubble's exterior interact with Earth's magnetic field to assist in controlling the angular momentum of the reaction wheels. The new pointing orientation is then determined by mapping a known field of stars using three star trackers. Finally, any two of three fine guidance sensors lock onto previously identified guide stars in the field of stars around the selected target. The fine guidance sensors are extraordinarily precise optical interferometers, capable of measuring the relative separations of stars in the sky to an accuracy of 0.002 arcseconds. They provide the ultimate degree of pointing stability to the telescope.

Commands to operate Hubble's suite of scientific instruments and digital data acquired from scientific observations with those instruments are routed through a command and data handling computer devoted to this purpose. Approximately 1.5 gigabytes per day of observational data are recorded onto either of two solid-state data recorders. The data are relayed through the spacecraft's two high-gain antennae to tracking and data relay satellites in geosynchronous orbit and from there to the TDRSS ground station at White Sands, New Mexico. The rate of production of scientific data from Hubble is expected roughly to double after the installation of new scientific instruments during the final servicing mission currently scheduled for 2005.

Hubble's Scientific Instruments

The HST can accommodate up to five scientific instruments. In addition, one of the three fine guidance sensors is designated as the primary instrument for astrometry—the science of precise measurement of the positions, motions, and distances of stars. NASA selected the first set of five Hubble instruments in 1977, and their designs reflected the technology of that era. The first set of instruments included a wide field and planetary camera, a faint object camera, a high-resolution spectrograph, a faint object spectrograph, and a high-speed photometer (25,26).

Hubble's designers intended that it be serviced and upgraded periodically; astronauts would convey new components to orbit and install them during space walks (EVAs). Regular service calls would allow the instrument technology to remain current. Thus, it was envisioned that Hubble's scientific performance would evolve steadily, as technology permitted and as the progress of science demanded. Consequently, when the telescope's spherical aberration was diagnosed

in 1990, NASA was well prepared to respond. All of the original instruments were affected by the poorly focused telescope image, some more adversely than others. In the first servicing mission in 1993, the wide field and planetary camera was replaced by an optically corrected instrument of similar design, which had been under development for several years before Hubble's launch. The high-speed photometer was removed from the observatory and replaced by COSTAR, the system of corrective optics that provided properly focused images to the two spectrographs and to the faint object camera. For the first time, these modifications allowed Hubble to achieve the level of scientific performance originally expected of it (27,28). Replacement of instruments on subsequent servicing missions, in each case, resulted in order-of-magnitude gains in scientific performance. In effect, each servicing visit in which an instrument was replaced has left a "new," far more capable observatory in orbit.

In general, Hubble's instruments fall into two categories—cameras and spectrographs. The cameras acquire images taken through a variety of selectable filters chosen by astronomers to transmit light in a particular range of colors or wavelengths. They provide insight into the structure, brightness, color, and distance of celestial objects ranging from neighboring planets in our own solar system to clumps of protogalaxies so far away that the light we record was emitted when the Universe was less than 10% of its present age. Spectrographs provide data that, while far less pleasing to the eye, are centrally important to studying the physical properties of planets, stars, nebulae, and galaxies. Spectrographic observations allow astronomers to measure the temperature, density, velocity through space, velocity of rotation, and chemical composition of objects that either emit or absorb light. Taken together, these two classes of instruments provide a powerful toolbox for exploring the Universe.

The evolution of the capabilities of Hubble's scientific instruments since 1993 reflects the technological evolution of light-sensing detectors. Foremost are the solid-state silicon detectors called charge-coupled devices (CCDs). Such sensors are commonly found in commercial products readily available to consumers—digital still cameras and video cameras. The CCDs used in HST instruments differ from their commercial cousins only in their level of performance; they must be highly sensitive and have fine spatial resolution, wide dynamic range, and very low background noise. The HST mission pioneered the application of CCDs to astronomical imaging by flying an early generation of the detectors into space in the first wide field and planetary camera. The second wide field and planetary camera (WFPC2, pronounced "wiff-pik two"), launched in 1993, incorporated somewhat improved CCDs. In analogy to current consumer products, one can describe WFPC2 as a 2-megapixel digital camera. The advanced camera for surveys (ACS) carried to Hubble in the fourth servicing mission in March 2002 supplanted WFPC2 scientifically. The ACS is a 16-megapixel digital camera that has twice the field of view of the sky, twice the angular resolution, and five times greater sensitivity than WFPC2—dramatic improvements in scientific capability enabled by advances in CCD technology (Figure 3).

Both the WFPC2 and the ACS are sensitive to wavelengths of light shorter than about 1000 nm (1 μm). They are intended to view the Universe in the colors of the spectrum to which the human eye is sensitive and also at shorter ultraviolet wavelengths down to about 115 nm. However, much important research



Figure 3. Hubble's new advanced camera for surveys acquired this image of the "Tadpole" galaxy shortly after it was launched in 2002. The long tail is the remnant of a collision between two galaxies, a small galaxy that has now been absorbed into the central region of a larger galaxy. This image astonished astronomers because of the great depth out to a large number of very distant galaxies it achieved in a small amount of observing time. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

depends upon being able to observe at wavelengths longer than $1\text{ }\mu\text{m}$ in the range of colors that are "redder than red," called the near-infrared. Light emitted by galaxies at the most distant reaches of the Universe is "red-shifted" by the expansion of the Universe from visible wavelengths into the near-infrared band. Red and near-infrared light are also not as efficiently absorbed or scattered by particles of dust as light of shorter wavelength—yellow, green, or blue (this is why sunsets look redder when there is a lot of dust in the air). The universe is filled with dust, much of it so thick that human vision cannot penetrate, for example, the dusty cocoons in which new stars are born. The HST was designed to provide access to near-infrared light emitted by celestial sources, particularly in the $1\text{--}2.5\text{ }\mu\text{m}$ wavelength range.

Humans sense infrared radiation with their skin as heat rather than as light seen with the eye. Sensors used to detect these wavelengths are very sensitive to heat and must be cooled to extremely low temperatures and carefully baffled to

avoid swamping the photons emitted by celestial sources by those emitted as heat by the warm telescope and its surrounding structures. Hubble's first infrared camera was the NICMOS (near-infrared camera and multi-object spectrometer), inserted into the spacecraft on the second servicing mission in February 1997. Its solid-state gallium arsenide (GaAs) detectors, as well as its filters and a portion of its optics, were encased in a 108-kg block of solid nitrogen ice. The ice cooled the detectors to an operating temperature of 60 K (-213°C). Although NICMOS was designed to operate in this fashion for approximately 4 years, a small thermal leakage introduced excess heat into the solid nitrogen, which sublimated after about 23 months in orbit. An accelerated observing program acquired most of the astronomical observations originally intended for NICMOS (29). However, researchers, who hoped to follow up their original observations or to pursue new research programs with NICMOS, sorely felt the premature loss of Hubble's near-infrared "eyesight." Hubble Project engineers and scientists found an ingenious solution to the problem by adapting a high-technology mechanical cooler, developed jointly by NASA and the U.S. Air Force to the task of cooling the dormant NICMOS instrument back to operating temperature. At the heart of this "reverse Brayton cycle" cryocooler are miniature turbines, approximately the diameter of a quarter in length, magnetically spun to very high rates, up to 7300 revolutions per second. This NICMOS cooling system (NCS) was installed during the servicing mission of 2002. It successfully resuscitated the instrument to full operation and provided improved instrument performance because it allowed engineers to set the temperature of the GaAs detectors at a more optimal value (77 K) than possible by using the solid nitrogen cryogen.

Two of Hubble's original instruments were spectrographs designed to operate in complementary fashion. The faint object spectrograph (FOS) concentrated on relatively faint targets observed with low spectral resolution; the Goddard high-resolution spectrograph (GHRS) provided higher resolution spectra of brighter sources. (Spectral resolution simply describes how finely light spread out into its component colors is sub-divided into measurable increments of color or wavelength.) Both instruments used similar light-sensing detectors developed in the 1970s called "digicons"—one-dimensional arrays of about 500 individual silicon diodes. Although these were powerful detectors in their day, they significantly constrained the capabilities of these early instruments. The FOS and GHRS were limited to observing a single point at a time in the sky—a single star or small patch of an extended source such as a planet or galaxy. Neither could efficiently acquire data that spanned a wide band of wavelengths. In the 1980s, the advent of highly efficient two-dimensional detector arrays containing millions of pixels provided the opportunity to advance Hubble's spectroscopic power dramatically. In 1997, on the second HST servicing mission, astronauts replaced both the FOS and GHRS with a single, far more capable instrument—the Space Telescope imaging spectrograph (STIS). The three detectors in STIS—one CCD and two electronic sensors incorporating microchannel plate technology—together span the wavelength range from approximately 110 nm to $1\text{ }\mu\text{m}$. For the first time on Hubble, a long entrance slit could be placed over an extended source, for example, across the nucleus of a galaxy, and the spectrum of each of 500 separate spatial points along the slit could be acquired simultaneously. This spatial and spectral multiplexing capability makes the STIS a

prodigious hunter of supermassive black holes in galactic nuclei, for example. In a given exposure time, STIS can also acquire data across a wider range of wavelengths at a given spectral resolution—by as much as a factor of about 30—than either of its predecessors. In short, STIS is a thoroughly modern scientific instrument that well illustrates the potential to enhance the power of Hubble by modernizing its technology during in-orbit servicing (29).

Two new instruments are being developed as future additions to the HST observatory. The wide field camera 3 (WFC3) will replace WFPC2. It is designed as a “panchromatic” camera, encompassing both ultraviolet/optical and near-infrared imaging capabilities in a single instrument. The WFC3 complements the ACS by providing greater sensitivity and field of view in the near-ultraviolet (200–300 nm) and provides a backup to the ACS between 300 nm and 1 μm . In the near-infrared from 1 to 1.7 μm , WFC3 will supersede the NICMOS and has a substantially larger field of view, higher angular resolution, and far greater sensitivity. The cosmic origins spectrograph (COS) is designed to be the most sensitive spectrograph ever flown in space. Although, like FOS and GHRS, it is intended primarily to observe point sources of light, these include extremely faint objects far across the cosmos, for example, distant quasars.

Servicing the HST

The Hubble Space Telescope observatory represents an intersection of the Human Space Flight program with Robotic Space Flight. In this regard, it is unique among NASA programs. Although the financial cost of human involvement is relatively high, the rewards are demonstrably great. Hubble is arguably NASA's most successful space science program. It has given humanity a realistic opportunity to seek answers to ancient and far-reaching questions: How did the universe begin? How did it come to look the way it does? What is the universe made of? How has it changed with time? Where did we come from? What is our destiny? Hubble has also provided humanity's first truly clear view of the beauty of the cosmos, providing inspiration and aesthetic satisfaction to people from all walks of life. Without the intervention of space-walking astronauts, Hubble's original optical flaw would not have been repaired, and the mission would ultimately have been judged a scientific failure. Without regular human servicing, Hubble would have ceased operation long ago. Without human servicing, Hubble could, ever have been renewed technologically, and its original capabilities would now be viewed as archaic. So, the cost of servicing Hubble must be weighed against the total cost of a hypothetical program of multiple replacements of a failed or archaic observatory that had comparable capabilities, for which servicing by humans was not an option.

From the beginning, Hubble was designed to be “human-rated,” modular, and serviceable (Figure 4). Out of a total of 51 major electrical, mechanical, and optical subsystems, only two—the primary telescope optics and the mechanical actuators that control their position or shape—are impossible for astronauts to replace or repair. Four fully successful Hubble servicing missions (as of 2002), encompassing a total of 18 fully successful space walks, have yielded a high level of experience and skill among the project engineers who design the missions and



Figure 4. Astronauts John Grunsfeld and Richard Linnehan install a new exterior radiator, part of the NICMOS Cooling System, on the Hubble Space Telescope during Servicing Mission 3B in March 2002. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the astronauts who execute them (30). The pinnacle of this capability was reached in the fourth servicing mission in March 2002, when astronauts John Grunsfeld and Richard Linnehan successfully replaced a failing central power control unit on Hubble despite the fact that this heavily cabled box, the size of a small bookcase, was not designed to be easily removed from the spacecraft.

The Space Shuttle crew assigned to Hubble servicing missions consists of seven astronauts—a mission commander, a Shuttle pilot, an operator of the Shuttle's remote manipulator robot arm (RMS), and a team of four space-walking (EVA) astronauts. Typically, the EVA team is assigned 15–20 months before each mission. This allows sufficient time for thorough familiarization with Hubble, EVA procedure development, and extensive training. Accurate replicas of the Hubble spacecraft structures are situated in the Neutral Buoyancy Laboratory (NBL), a 40-foot deep, 6.2-million gallon swimming pool at the Johnson Space Center in Houston, Texas. There the astronauts can practice the choreography of every task of every space walk in an environment where the buoyancy of water simulates the weightless state they will encounter in space. Extensive crew training is also conducted at the Goddard Space Flight Center in Greenbelt, Maryland, where the crew practices removing and inserting the actual flight hardware that they will carry to orbit into high fidelity mechanical, electrical,

and computer simulators of the Hubble spacecraft. The flight hardware is stored on carriers and within protective boxes in the Shuttle payload bay. The astronauts learn to remove the new flight components, including scientific instruments, from their carriers and to stow hardware removed from Hubble for the return journey to Earth. A large assortment of crew aids and tools—lights, tethers, handling brackets, foot restraints, and numerous manual and power tools—are developed to facilitate the EVA tasks and to minimize the time required and the physical stresses placed on the EVA astronauts. The training process is iterative in that, as the astronauts practice each task in the NBL or at Goddard, they often devise alternative procedures or recommend modifications of the tools and other hardware to improve the chances that each task can be successfully completed. Their ideas then become the basis for further improvements.

During each servicing mission, Hubble remains (with a few exceptions) powered on and operating even while physically attached to the Shuttle. Each EVA task must be coordinated with spacecraft operators at the Space Telescope Operations Control Center at Goddard, who also must be prepared to respond quickly and effectively to any unforeseen problems that arise as each EVA and the overall mission progresses. Before each mission, the teams at Goddard and Johnson, including the astronauts, conduct numerous dress rehearsals to practice every step of the flight. These joint integrated simulations involve practicing responses to plausible mission contingencies and emergencies that are often more dire than those typically experienced during an actual flight. In this way, the entire flight and ground-based operations team becomes skilled at complex problem solving—a major contributor to the apparent smoothness with which the actual missions are conducted.

A typical Hubble servicing mission lasts approximately 11 days, commencing with a spectacular launch of the Shuttle from Launch Complex 39A or 39B at the Kennedy Space Center in Florida. During the first two days of flight, the Shuttle crew initiates a number of thruster “burns” to adjust its orbit and to catch up gradually to the HST. On the third flight day, a final set of maneuvers brings the Shuttle within approximately 35 feet of the telescope. At the same time, ground controllers command HST to rotate so that its two grapple fixtures are properly oriented for capture by the Shuttle’s robot arm. Once captured by the arm operator, Hubble is berthed and latched to a platform in the payload bay, somewhat akin to a “lazy Susan,” which can be commanded to pivot and rotate to bring the various work sites on the spacecraft within easy reach of the EVA astronauts.

The core of each mission is the sequence of four or five EVAs carried out on successive days. Two EVA teams, each comprised of two astronauts, conduct their space walks on alternating days. The planned duration of each EVA is approximately 6 hours. However, it is not unusual for Hubble-servicing EVAs to extend as long as 8 hours. The ultimate limitations on EVA duration are set by spacesuit oxygen supplies, the degree of fatigue of the EVA team, and the time necessary to close down the work area and return to the airlock. During each EVA, one astronaut works from the end of the robot arm, attached to the RMS by foot restraints. The other astronaut works as a “free-floater,” maneuvering around the payload bay with the assistance of tethers and handholds. The two exchange positions as the EVA progresses. The RMS operator inside the Shuttle

cabin is a critical third member of the EVA team. He or she positions the astronaut attached to the arm with fine precision to allow opening Hubble's compartment doors, unplugging an instrument or other module, removing and stowing it elsewhere in the Shuttle payload bay, extracting new equipment from its protective enclosure or carrier and installing it within Hubble, followed by stowage of the old module for the return home. One unplanned EVA day is set aside in case it is needed to complete any unfinished tasks, and another unplanned EVA day is available to attend to any contingencies for final deployment of HST back into orbit or for preparing the Shuttle itself for reentry and landing. For example, if Hubble's aperture door should fail to open, astronauts could execute a "contingency EVA" to crank it open manually. The final 2–3 days of the flight provide an opportunity for the Shuttle crew to rest, to conduct other tasks unrelated to Hubble, and to prepare the orbiter for the return home.

To date (2002), four Hubble servicing missions have been completed. Servicing Mission 1 (SM1) in December 1993 was primarily oriented to correcting the telescope's optical flaw by installing WFPC2 and COSTAR. However, the astronauts also installed a new set of solar arrays, solar array drive electronics, new gyros, and additional memory and processing power for the spacecraft's central computer. They also made a simple repair to the Goddard high-resolution spectrograph.

In February 1997, Servicing Mission 2 (SM2) provided the first opportunity to concentrate on upgrading Hubble rather than simply repairing it. In this mission, two new, advanced technology instruments—the Space Telescope imaging spectrograph (STIS) and the near-infrared camera and multi-object spectrometer (NICMOS) were installed in the instrument bays from which the two first-generation spectrographs (FOS and GHRS) were removed. In addition, a new, solid-state digital data recorder replaced one of Hubble's old mechanical tape recorders, and a failing fine guidance sensor was replaced by a spare unit so that the former could be returned to the ground for refurbishing. Other replacements included a reaction wheel, a data interface unit, and solar array drive electronics. On this mission, the astronauts saw, for the first time, signs that the external thermal blankets covering Hubble's exterior were beginning to crack and peel. They reacted to this contingency, guided by engineers on the ground, by improvising some temporary thermal covers to patch the more severely degraded areas.

What had originally been planned as a single Servicing Mission 3 was split into two separate missions (SM3A and SM3B) when it was realized early in 1998 that Hubble's gyros were failing at an alarming rate. Mission 3A was quickly planned as a "contingency mission" to replace all of the spacecraft's six gyros. Three gyros must be working for Hubble to conduct science operations. Approximately six weeks before the launch of SM3A in December 1999, the fourth gyro failed, science observations by Hubble ceased, and the spacecraft was placed into a "zero-gyro" safemode until the astronauts arrived to make the repair. Advantage was taken of this unforeseen servicing opportunity to place a new, more capable central computer into the HST, to add a second digital data recorder, to replace a faulty radio transmitter, and to continue the round-robin change-out and refurbishing of the fine guidance sensors.

The most complex and difficult servicing mission to date was SM3B, launched in March 2002. The mission dramatically upgraded Hubble's scientific

capabilities by the insertion of the advanced camera for surveys (ACS) and the NICMOS cooling system (NCS). In addition, the astronauts began the process of completely overhauling and updating the spacecraft's electrical power system by installing new, rigid, high-powered solar arrays and replacing the aging power control unit (PCU). Although the total list of EVA tasks on SM3B was shorter than that on prior missions, most of those tasks were more difficult and required much more EVA time to execute than on the previous three missions. SM3B was probably at the limit of what space-suited astronauts can physically accomplish, and this particular crew acquitted itself with glory.

At this time, a final servicing mission to Hubble is planned for 2005. Two new instruments (WFC3 and COS) will be installed, and a new thermal radiator will be mounted on Hubble's exterior to help keep the scientific instruments cooler during the latter years of the mission. The refurbishing and change-out of all three fine guidance sensors will be completed, the electrical overhaul of HST will be completed by replacing all six of its NiH_2 batteries, and a new set of six gyros will be installed. When the astronauts leave Hubble in orbit for the last time at the end of SM4, it will be at the apex of its capabilities—more powerful scientifically and more modern and robust technologically than at any time in its history. NASA currently plans to retrieve Hubble by using the Shuttle and to return it to the ground sometime around 2010.

Operating Hubble

Two categories of operations are necessary for Hubble, as for any space observatory—Science Operations and Spacecraft Operations. Science Operations translate the research plans of scientists into detailed sequences of specific commands to the internal electronics, detectors, and mechanisms of the scientific instruments and coordinates the execution of astronomical observations with the operations of the spacecraft. This coordination includes selecting of guide stars on which the fine guidance sensors will lock to hold the telescope steady during observations. It also specifies the timing of observations, to take into account when targets are blocked by Earth during each 94-minute Hubble orbit and cannot be seen and when passage of Hubble through the most intense region of the van Allen radiation belts (the South Atlantic Anomaly) would degrade the observations. Spacecraft Operations control the functions of the spacecraft itself—loading commands into the onboard computers; pointing Hubble to the desired position on the sky and locking onto the guide stars; collecting and routing electrical power from the solar arrays; and storing digital data from the instruments, routing the data to onboard radio transmitters, and relaying the data to the Tracking and Data Relay Satellite System (TDRSS) for transmission to the ground. Spacecraft Operations also monitor all of the systems on Hubble to make sure that they are always operating properly, that Hubble remains “healthy” and “safe.” If not, ground controllers intervene to remedy the problem, or in more serious contingencies, the spacecraft places itself into a protective “safe mode” to wait for troubleshooting and intervention from the ground. Both types of operations—Science and Spacecraft—must be seamlessly blended together 24 hours per day, every day of the year. The Space Telescope Science Institute (STScI) in Baltimore, Maryland, manages Hubble's Science

Operations. The Hubble Operations Project at the Goddard Space Flight Center in Greenbelt, Maryland, is responsible for Spacecraft Operations, including all activities associated with Hubble's health and safety.

The Hubble observatory is operated as a "public facility." Once per year, astronomers from all over the world are invited to submit observing proposals for research that requires Hubble's unique capabilities. Typically, the amount of observing time requested on the telescope exceeds the total available time by factors of 6–8. Competition to use Hubble is intense, and many truly excellent scientific research proposals must be turned down each year. All proposals are peer-reviewed by panels of independent scientific experts who recommend to the Director of the Space Telescope Science Institute those proposals that have the most compelling scientific merit. The astronomers, whose proposals are selected, then provide very detailed descriptions of what they want the telescope to do—targets to be observed, instrument modes to be used, exposure times required, etc. A computer program at the STScI ingests these requests from all observers and produces a tentative calendar that extends months into the future, attempting to optimize the overall observing efficiency of the observatory. Celestial objects observed from Earth are frequently available only during particular seasons of the year. So, for example, the long-range planning calendar must take target availability into account. The detailed commanding plan for Hubble is laid out weekly. Approximately 11 days before the beginning of a specific week, the STScI delivers computer output to the spacecraft operators at Goddard that contains the detailed specifications and time line of commanded events for the science observations of that week. This "science mission specification" is then integrated into the overall spacecraft command plan for the week. During the week, the resulting spacecraft and instrument commands are uplinked and stored on board Hubble's computers several times per day. On average, about 15,000 stored commands are sent to Hubble each day. Direct control of Hubble from the ground to execute science observations is rare. The spacecraft is fully robotic and executes the preplanned science program autonomously under internal computer control.

During a single week, about 650 individual exposures on celestial targets are acquired. In October 2002, Hubble completed its 500,000th exposure. Currently, Hubble returns an average of 45 gigabytes of data to the ground each month. The data are received at Goddard, re-formatted and transmitted to the STScI. There, the data are processed, calibrated, archived, and distributed to the scientists who originally requested them. The scientists have a 1-year period during which their observations remain proprietary, so that they may be fully analyzed and interpreted. The results are usually published in professional scientific journals. After the 1-year proprietary period expires, the data become easily accessible to the remainder of the scientific community and to the public by direct access via the World Wide Web to the Hubble data archives at the STScI. Currently, the Hubble archives hold approximately 10 terabytes of scientific data, most of which are in the public domain.

Hubble's Scientific Achievements

The Hubble Space Telescope does not work in isolation. It is the flagship of a growing fleet of modern astronomical telescopes in space and on the ground. The

unique power of the HST derives from its combination of extremely sharp images that cover relatively wide angular fields in the sky and have a deep dynamic range, low background noise, and sensitivity to wavelengths from the vacuum ultraviolet to the near-infrared. Most of Hubble's accomplishments build upon the previous work of ground-based and space astronomers over many decades. Hubble's greatest achievement is its facility for converting so many prior hypotheses, for which supporting empirical data were scant, ambiguous, and painfully difficult to obtain, into clearly and decisively demonstrated truth. But the HST has gone well beyond that. It has provided a detailed view of the unimagined complexity and diversity of the universe, as well as its startling beauty (31–34). It has yielded numerous surprises and raised new questions. With each new instrument inserted by the astronauts on servicing missions, Hubble grows in capability by factors of 10. It can reasonably be anticipated that Hubble's second decade will be at least as fruitful scientifically as its first. Some of the most important ways in which Hubble has influenced scientific thought follow.

Imaging the Distant Universe. The HST provided the first deep, clear view of the distant Universe from the approximately 150-hour exposure on the northern Hubble Deep Field in 1995 (35,36). Only the COBE satellite has probed farther back in time, measuring the radiation left over from the Big Bang itself. The HST has shown that, when the Universe was very young, it was populated by structures that were much smaller and much more irregular in shape than the galaxies that we see in the modern Universe. It is believed that these smaller structures, made up of young stars and primordial gas, are the building blocks from which the more familiar spiral and elliptical galaxies were formed. However, the processes were complex, involving multiple galaxy collisions, in-fall of intergalactic gas, and the gravitational influence of supermassive black holes.

Precise Calibration of the Distance Scale. The HST was the first telescope capable of resolving the “standard candle” Cepheid variable stars and using them to obtain very accurate distances to a large number of moderately distant galaxies. These distances were used in turn to recalibrate a number of other standard distance indicators such as Type Ia supernovae, which were applied in extending distance measurements to galaxies at much greater distances out into the “Hubble flow” where the relative velocities of galaxies are dominated by the expansion of the Universe itself. The result is a much more accurate measure of the rate at which the Universe is expanding (the Hubble constant) and a determination of the age of the Universe, that is, how long it had to have been expanding for galaxies to have reached the presently measured distances separating them (37). The initial calculation of the age of the Universe from Hubble data was based on the assumption that the expansion of the Universe is slowing down under the pull of the gravity exerted by all of the mass within it. In other words, the Universe was expanding more rapidly in the past than it is today. On this basis, the calculated age was 9–11 billion years, a value less than the ages independently estimated, for example, of the oldest stars in our own galaxy. This seemingly absurd contradiction was resolved only when it was determined that the expansion of the Universe is accelerating, rather than decelerating, at the present time, as discussed below.

Measuring the Cosmological Constant and the Age of the Universe. In its first decade, the HST partnered with ground-based telescopes in searching for

and measuring the peak brightness and rate of change of brightness (“light curves”) of Type Ia supernovae in distant galaxies whose light was emitted when the Universe was about half its present age (redshifts up to $Z = 1.2$). Hubble’s major contribution was accurate measurement of the most distant supernovae in this sample. From these measurements, the galaxies’ distances could be accurately determined, and these values, combined with the measured recessional velocities of the galaxies, indicated the rate at which the Universe itself was expanding in various epochs far back in time. The result was remarkable and provided the first clue that the expansion of the Universe is accelerating at the present time—driven by an unknown repulsive form of gravity named “dark energy.” Einstein anticipated such a possibility by adding a “cosmological constant” to his equations of general relativity because he believed that the Universe had to be static and would collapse under its own gravity if there were no repulsive force to compensate for ordinary gravity. After learning from Edwin Hubble that the Universe is observed to be expanding and thus, is not static, Einstein removed the cosmological constant from his theory, referring to it as his “greatest blunder.” In 2001, the HST set the record for the most distant Type Ia supernova ever discovered (38). This supernova exploded when the universe was only one-third of its present age, and the HST’s measurements show that the expansion of the Universe was still decelerating under the influence of ordinary gravity in that epoch. This discovery placed the evidence for the relatively recent transition of the expansion from deceleration to acceleration on far firmer ground by eliminating several alternative interpretations of the earlier observations of less distant supernovae, such as the possibility that their brightness is dimmed by intervening dust. If the Universe is currently accelerating, it took longer to reach its present size, and so is older than one would calculate from a simple measurement of the current value of the Hubble constant. The calculated age, properly corrected for acceleration, is 13–15 billion years—consistent with the ages estimated for the oldest stars (39). Currently, the accelerating universe and the nature of “dark energy” are considered among the most important and baffling problems in modern physics. Much future research must be done with the Hubble and other telescopes to “tease” clues about it out of the fabric of the Universe. The pursuit of “dark energy” may well produce a revolution in our understanding of the fundamental laws of physics.

Detecting and Measuring Supermassive Black Holes. Before the launch of the HST, ground-based images of galactic nuclei hinted at the existence of large concentrations of mass at the very centers of galaxies. Although it was suspected that these might be the massive black holes predicted theoretically as early as the 1930s, this was impossible to prove at the resolution of ground-based optical telescopes. The HST was the first optical telescope capable of probing sufficiently close to the center of a galaxy to measure spectroscopically the velocity of stars and gas in orbit around the central concentration and to measure accurately by directly imaging the size of the central cusp of starlight. Thus, in 1994, Hubble provided conclusive proof that a central black hole several billion times the mass of our Sun exists at the core of the giant elliptical galaxy M87. At about the same time, a ground-based microwave telescope measured the velocity of water masers in orbit around a black hole of several million solar masses in a different galaxy, thus providing further proof. The HST has now moved beyond

the initial confirmation of the existence of supermassive black holes to a “demographic” survey of central black holes. Hubble has demonstrated that these powerful, enigmatic objects are found in the nuclei of most (or perhaps all) galaxies, whether or not those nuclei are energetically active. More profoundly, Hubble observations show that there is a very tight correlation between the mass of the central black hole in a galaxy and the mass of the surrounding ellipsoidal “bulge” of ordinary stars in which the black hole resides, whether observed in an elliptical galaxy or within the central bulge of a spiral galaxy like our own Milky Way. This relation is observed across the full range of black hole masses from one million to several billion times the mass of the Sun (40). Recently, Hubble observations provided the first evidence of possible intermediate-mass black holes thousands of times the mass of the Sun at the center of another type of ellipsoidal (spherical) system of stars—the globular star clusters that populate our own and other galaxies. The preliminary results indicate that these intermediate-mass black holes obey the very same relationship to the mass of the system of stars in which they exist as the supermassive black holes at the centers of galaxies. Mother Nature is providing a very strong clue here about the relationship between the formation of black holes and the formation of galaxies and other star systems—a clue that has yet to be deciphered.

The Nature of Quasars. For several decades, quasars (quasi-stellar radio sources), sometimes called QSOs (quasi-stellar objects), were among the most enigmatic objects in the Universe. When they were discovered in the 1960s, they were recognized as the most distant and energetic objects known. Continued study suggested a possible relationship between the quasars and another puzzling phenomenon—the highly active and energetic nuclei of certain galaxies at more moderate distances, the AGN (for active galactic nuclei). The detection of very faint “fuzz” around some quasars, seen with ground-based telescopes, supported the hypothesis that they might be very distant AGNs in the early Universe that are undergoing especially intense outbursts of activity. The HST has completely verified this idea (41). The telescope’s resolution and dynamic range clearly reveal a variety of underlying host galaxies of quasars. A more surprising HST discovery is that a large fraction of quasar host galaxies appears to be in the process of colliding and merging with other galaxies. This suggests that galaxy collisions, which the HST has shown are common in the early Universe, may have provided the extra “fuel” to the massive central black holes in galaxies that is needed to generate the prodigious energy output of quasars and AGN. We know now that most or all galaxies possess massive central black holes, so it is assumed that those galaxies that are “quiet” at their nuclei, such as our own Milky Way, must be in a quiescent state, lacking a source of fuel, namely in-falling stars and gas.

The origin of Gamma-ray Bursts. Intense bursts of highly energetic gamma radiation from unknown cosmic sources were first detected by military satellites. Thousands of these bursts were subsequently observed by the Compton Gamma Ray Observatory, which found that they are distributed more or less uniformly across the sky. Not only was the source of the bursts a mystery, it was not even known if they originated in our own galaxy, far across the Universe, or somewhere in between. The joint Italian-Dutch satellite Beppo-Sax was designed to spot gamma-ray bursts very quickly and to locate their positions accurately, so that other telescopes could be trained on them while they were still bright. Using

such “alerts” from Beppo-Sax, ground-based telescopes then located the gamma-ray sources in optical-wavelength light. Using this information, astronomers trained the HST on the optical counterparts of multiple gamma-ray bursts (42). Hubble’s resolution and sensitivity gave it the unique ability to show that the sources of the gamma-ray bursts are embedded in faint, distant galaxies. Moreover, the optical light from the fading bursts in most cases emerges at random distances from the centers of these galaxies and is usually associated with regions undergoing episodes of intense massive star formation. The bursts do not appear to be associated with central black holes and active galactic nuclei. By following the brightness changes in the sources to very faint levels, the HST provides important constraints on models of the stellar “catastrophes” that produce these extraordinarily intense and rapid outbursts of energy. Today, at least two alternative explanations have been offered to explain the origin of gamma-ray bursts. Either an extremely massive star explodes, producing what has been dubbed a “hypernova” (implying explosive energy release far exceeding that of more common supernovae and perhaps concentrated in beams of radiation), or perhaps two neutron stars collide, forming a black hole. In the future, astronomers will attempt to select from among such alternative explanations on the basis of more detailed observations of the environments of gamma-ray bursts and also using measurements of the bursts themselves obtained closer to the moment of their peak brightness. The High Energy Transient Explorer (HETE-2) satellite, launched in 2001, now facilitates more rapid responses by Hubble and other telescopes to gamma-ray bursts.

The Birth of Stars. The HST’s resolution and sensitivity to both visible and infrared light have given it unprecedented, clear views of the rich, diverse, and complex processes that lead to star formation. The collision of two galaxies, as clearly seen by the HST, stimulates the births of large populations of young, massive stars and star clusters (43). Compression of interstellar gas by the intense radiation from a massive star can trigger the formation of smaller stars nearby. It is seen that the radiation and ejected material from supernova explosions compress and enrich the interstellar gas and dust from which new stars can form. Stars forming in large, dense clouds of molecular hydrogen and dust are limited in the masses they can achieve by the erosion of material from those clouds by radiation from nearby hot stars. The formation of an individual star always seems to be governed by an accretion disk of material falling onto the young protostar and by highly aligned bipolar jets carrying material away from the “construction site.” All of these processes have been clearly elucidated by Hubble observations (44).

The Formation of Planetary Systems. Before the HST, the presence of dust disks around a small number of young stars had been inferred from observations by infrared satellites, and a ground-based coronagraphic instrument had directly imaged one such disk around the star, beta Pictoris. For centuries, it has been believed that such a disk must have been the precursor to our own solar system, providing the raw material from which the planets were constructed. The existence of protoplanetary disks around other stars is, therefore, a necessary condition for the existence of extrasolar planetary systems. The HST has revolutionized this area of science. Images of the Orion nebula region obtained by Hubble (45) revealed that a large proportion of young stars (about 50%) is



Figure 5. Three “proplyds” (for “protoplanetary disk”)—young stars surrounded by disks of dust and gas from which a system of planets might form—seen in the Orion Nebula. The “cocoon” shaped structures are envelopes of gas being blown out of the disk by the intense radiation from a nearby hot, young star. The disk seen in silhouette toward the upper left is apparently shaded from the hostile radiation environment and stands a better chance of forming a planetary system. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

surrounded by gas and dust structures, many of which are clearly disks (Figure 5). High-resolution Hubble near-infrared images penetrating the dense dust of the Taurus dark cloud and other star-forming regions show protoplanetary disks in the process of formation and evolution (46). These disks are common and they contain enough material to form entire planetary systems equivalent to our solar system. HST coronagraphic observations revealed, for the first time, the internal structures of protoplanetary disks and debris left behind by prior planet formation (47). Hubble has also revealed the difficulties many stars may have in forming planetary systems because of the hostile environments in which they are born. Observations of the Orion nebula, for example,

reveal that gas contained in protoplanetary disks may be quickly blown out of the disks by the intense radiation and fluxes of particles emitted by nearby massive, hot young stars. If a newly formed star the mass of our Sun, for example, is shaded by nearby clouds of dust, its surrounding protoplanetary disk will retain the hydrogen and helium gas needed to build gas-giant planets like Jupiter or Saturn. However, if the star is out in the open and exposed to the radiation of a nearby massive star, the formation of its gas-giant planets becomes a race against time—will new planets form before the necessary raw material has been blown away? The HST has opened this fertile new area of observational astronomy—the empirical study of the origin, structure, and evolution of protoplanetary systems.

The Death of Stars. Dying stars shed material into interstellar space, sometimes gently and episodically, sometimes in explosive catastrophes. In either case, the ejected material is enriched in chemical elements produced in the interior nuclear furnaces of these stars and thus “seeds” the interstellar gas and dust with the basic building blocks from which new stars, planets, and life may form. The HST has provided exquisite images of dying stars. These are the basis for a remarkably detailed understanding of the events preceding the deaths of stars—how material is shed from dying stars, how that material interacts with the environment around the star, and how the process is influenced by each star’s individual circumstances. Was the star single or part of a multiple star system, did it have planets, did it have a magnetic field, was it rapidly rotating, etc.? These factors determine the complex and incredibly beautiful structures of so-called “planetary nebulae”—the remnants of the outer layers of red giant stars which, having exhausted their thermonuclear fuel, become unstable and eject most of their mass into interstellar space. Only the star’s core remains, a hot white dwarf whose intense radiation is absorbed by the ejected material, causing that material to glow. Stars like the Sun (and a bit more massive than the Sun) end their lives in this manner. The gentle ejection of planetary nebulae from such stars is the primary source of carbon in the Universe, the basis of our organic chemistry. The most spectacular example of a massive dying star is Hubble’s imagery and spectroscopy of supernova SN1987a. For the first time, astronomers saw the delicate ring structures ejected during the preexplosive evolution of the dying star. They saw the blast debris expanding outward over time from the supernova explosion. Now they are seeing the innermost ring “light up” as the blast material plows into it (48).

Our Dynamic Solar System. The HST has obtained beautifully detailed images of the planets, satellites, comets, and asteroids of our own solar system (49) regularly for a period of years. It cannot rival the images taken by flyby spacecraft or orbiting probes like the Voyager or Galileo missions. However, HST complements these *in situ* missions because it can observe the entire solar system and follow changes across long periods of time by imaging and spectroscopy of exquisite quality. Hubble provided the first resolved images of Pluto and its satellite Charon, enabling measurement of their masses and crude mapping of their surfaces. HST imagery has shown that the atmospheres of the gas-giant outer planets, Uranus and Neptune, once thought to be bland and nearly featureless, possess very dynamic climates. Giant cloud patterns form and dissipate with regularity. The ultraviolet imaging capability of STIS and WFPC2 have given

planetary scientists remarkable views of the northern and southern lights on Jupiter, Saturn, and Ganymede. Using Hubble, scientists have traced the dynamic electrical interactions between Jupiter and its satellite Io. In 1995, astronomers had the rare opportunity to view Saturn's rings edge-on. Hubble's sharp resolution led to the discovery of a diffuse atmosphere surrounding the rings, the discovery of several new satellites, and the recovery of old satellites in strange positions. The implication is that we are watching satellites of Saturn being created and destroyed nearly in "real time." The HST has monitored the weather on Mars and has provided remarkable images of seasonal changes at the Martian poles. In 1994, Hubble obtained uniquely clear pictures of the collisions of the 21 fragments of Comet Shoemaker-Levy/9 with the upper atmosphere of Jupiter and their aftermath (50). These revealed the enormous fireballs created when fragments entered the Jovian atmosphere at 140,000 miles per hour and heated the atmospheric gases up to 50,000°F, cooking them into a stew of "soot" and organic molecules (Figure 6). Studies of atmospheric waves propagating from the impact sites gave unique new information about the composition and density of Jupiter's atmosphere. The dispersal of the "soot" during several weeks allowed scientists to monitor the upper atmospheric winds. But the greatest contribution of the observing campaigns on Comet S-L/9 by the HST and by many other telescopes on earth was to remind humanity of our vulnerability as a planet and to motivate us to remain vigilant to the space environment in which we exist.

The Unexpected. One measure of the impact of Hubble observations on scientific thought is the frequency with which scientific discoveries made using

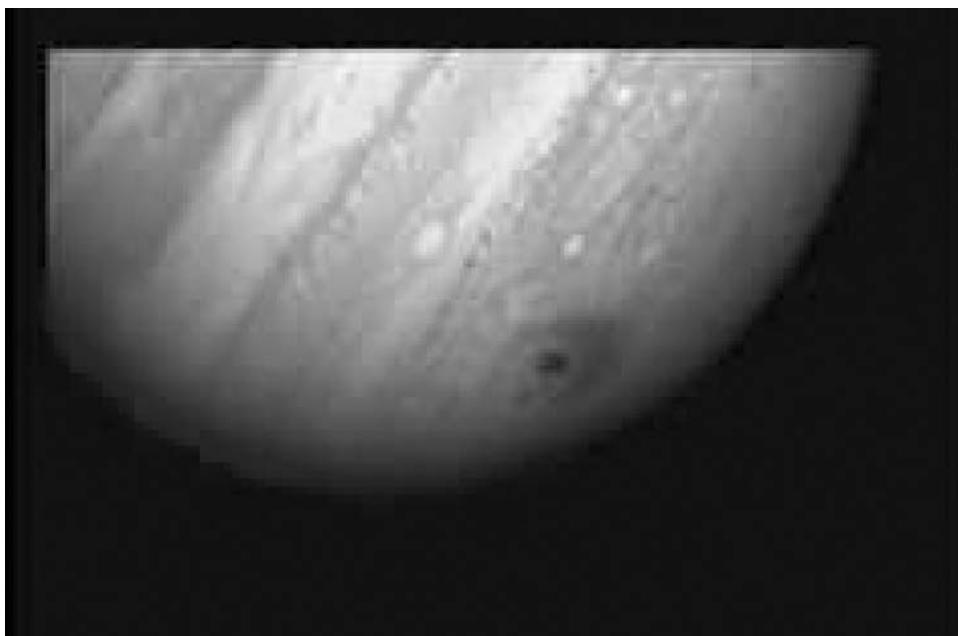


Figure 6. Impact zone of a large fragment of comet Shoemaker-Levy 9 in the atmosphere of Jupiter, as observed by Hubble in July 1994. The crescent pattern of "soot" floating high above Jupiter's cloud tops is roughly the diameter of Earth. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Hubble observations are cited by the authors of other papers in major scientific journals. In other words, how influential are the results from Hubble on the work of other scientists? Using this criterion, one can list in rank order those areas of research on which Hubble has made the greatest positive impact. Four of the top ten areas—galactic evolution (particularly from the Hubble Deep Field observations), the accelerating Universe, gamma-ray bursts, and dust disks around young stars—were either not known or were not expected to be subjects for major Hubble discoveries before the telescope was launched in 1990. This illustrates the great virtue of having available a powerful, versatile space observatory, regularly updated, maintained, and at the disposal of astronomers to address the most compelling scientific problems of the day, many of which were unanticipated only a few years previously. It is reasonable to expect that the Hubble Space Telescope will continue to compel us to address questions that we once did not even know how to ask for the remainder of its lifetime in orbit.

BIBLIOGRAPHY

1. Petersen, C.C., and J.C. Brandt. *Hubble Vision: Astronomy with the Hubble Space Telescope*. Cambridge University Press, New York, 1995.
2. Christianson, G.E., *Edwin Hubble: Mariner of the Nebulae*. Farrar, Straus and Giroux, New York, 1995.
3. Spitzer, L., et al. *Scientific Uses of the Large Space Telescope*. National Academy of Sciences, Washington, DC, 1969.
4. Longair, M.S., and J.W. Warner (eds). *Scientific Research with the Space Telescope*, IAU Colloquium No. 54. U.S. Government Printing Office, Washington, DC, 1979.
5. Hornig, D., et al. *Institutional Arrangements for the Space Telescope*. National Academy of Sciences, Washington, DC, 1976.
6. Smith, R.W. *The Space Telescope: A Study of NASA, Science, Technology and Politics*. Cambridge University Press, New York, 1989.
7. Allen, L., et al. *The Hubble Space Telescope Optical System Failure Report*. NASA Publ. TM-103443, Washington, DC, 1990.
8. Burrows, C.J., et al. *Astrophys. J.* 369: L21–L25 (1991).
9. Wood, H.J., and S.W. Hinkal. *Optical Alignment IV*. SPIE Proceedings. 1996: 134–143 (1993).
10. Brown, R.A., and H.C. Ford (eds). *Report of the HST Strategy Panel: A Strategy for Recovery*. Space Telescope Science Institute, Baltimore, 1991.
11. Harnisch, R.J., and R.L. White (eds). *The Restoration of HST Images and Spectra-II*. Space Telescope Science Institute, Baltimore, 1993.
12. HST Special Issue. *Astrophys. J.* 369: L26–L78 (1991).
13. HST Special Issue. *Astrophys. J.* 377: L1–L64 (1991).
14. Benvenuti, P., and E.J. Schreier (eds). *Science with the Hubble Space Telescope*. European Southern Observatory, Garching, 1992.
15. Cowen, R. *Science News* 144: 296–298 (1993).
16. Fienberg, R.T. *Sky & Telescope* 86: 16–22 (1993).
17. Hoffman, J.A. *Sky & Telescope* 86: 23–29 (1993).
18. Cowen, R. *Science News* 145: 52 (1994).
19. Fienberg, R.T. *Sky & Telescope* 87: 20–27 (1994).
20. Hawley, S.A. *Sky & Telescope* 93: 42–47 (1997).
21. Fienberg, R.T. *Sky & Telescope* 93: 46 (1997).
22. Grunsfeld, J.M. *Sky & Telescope* 99: 36–42 (2000).

23. Grunsfeld, J.M. *Sky & Telescope* 103: 30–33 (2002).
24. Schroeder, D.J. *Astronomical Optics*. Academic Press, San Diego, 2000.
25. Leckrone, D.S. *Publ. Astron. Soc. Pacific* 92: 5–21 (1980).
26. Leckrone, D.S. *Philos. Trans. R. Soc. London* 307: 549–561 (1982).
27. HST Special Issue. *Astrophys. J.* 435: L1–L78 (1994).
28. Benvenuti, P., F.D. Macchetto, and E.J. Schreier (eds). *Science with the Hubble Space Telescope—II*. Space Telescope Science Institute, Baltimore, 1995.
29. HST Special Issue. *Astrophys. J.* 492: L83–L184 (1998).
30. Hubble Space Telescope Project. Available <http://www.hubble.gsfc.nasa.gov>
31. Voit, M. *Hubble Space Telescope: New Views of the Universe*. Abrams, New York, 2000.
32. Harwood, W. *Hubble: The Space Telescope's Eye on the Cosmos*. Pole Star, Kent, U.K., 2002.
33. Petersen, C.C., and J.C. Brandt. *Hubble Vision: Further Adventures with the Hubble Space Telescope*. Cambridge University Press, Cambridge, 1998.
34. Space Telescope Science Institute. Available <http://www.stsci.edu>
35. Ferguson, H.C., R.E. Williams, and L.L. Cowie. *Physics Today* 50: 24–30 (1997).
36. Livio, M., S.M. Fall, and P. Madau. *The Hubble Deep Field: Proceedings of the Space Telescope Science Institute Symposium*. Space Telescope Science Institute, Baltimore, 1997.
37. Freedman, W.L., et al. *Astrophys. J.* 553: 47–72 (2001).
38. Riess, A.G., et al. *Astrophys. J.* 560: 49–71 (2001).
39. Freedman, W.L. In *Astrophysical Ages and Time Scales*, Astron. Soc. Pacific Conf. Ser. 245: T. von Hippel et al. (eds), 2001, pp. 542–551.
40. Gebhardt, K., et al. *Astrophys. J.* 539: L13–L16 (2000).
41. Boyce, P.J., et al. *Mon. Not. R. Astron. Soc.* 298: 121–130 (1998).
42. van Paradijs, J., C. Kouveliotou, and R. Wijers. *Annu. Rev. Astron. Astrophys.* 38: 379–425 (2000).
43. Whitmore, B.C., et al. *Astron. J.* 118: 1551–1576 (1999).
44. Reipurth, B., and J. Bally. *Annu. Rev. Astron. Astrophys.* 39: 403–455 (2001).
45. O'Dell, C.R., *Annu. Rev. Astron. Astrophys.* 39: 99–136 (2001).
46. Padgett, D.L., et al. *Astron. J.* 117: 1490–1504 (1999).
47. Schneider, G., et al. *Astrophysical Ages and Time Scales*, Astron. Soc. Pacific Conf. Ser. 245: T. von Hippel et al. (eds), 2001, pp. 121–129.
48. Pun, C.S.J., et al. *Astrophys. J.* 572: 906–931 (2002).
49. James, P.B., and S.W. Lee. *Annu. Rev. Earth Planetary Sci.* 27: 115–148 (1999).
50. HST Special Issue on Comet Shoemaker-Levy 9. *Science* 267: 1237–1392 (1995).

DAVID S. LECKRONE
 NASA, Goddard Space Flight Center
 Greenbelt, Maryland

HUMAN OPERATIONS IN SPACE DURING THE SPACE SHUTTLE ERA

When the Space Shuttle Columbia was launched on its maiden flight 20 years and more than 100 flights ago, it carried a crew of two, an instrumentation pallet

as its only payload, and the key mission objectives were to see if launch and landing could be performed safely. That launch also carried the promise of a new era of space operations, including research in space, launching and retrieving of satellites, in-orbit maintenance of spacecraft, and ultimately the building and use of a space station. Slowly and deliberately during these 20 years, the flight team on the ground and in orbit has developed the maturity of operations in space and the goal of achieving complex operations with humans from the Space Shuttle. This article looks back at the last 20 years of Space Shuttle operations to review what has been accomplished in attaining these objectives. The goal of this review is not to serve as a chronology of all 100 flights, but rather to highlight some of the key milestones in developing the sophisticated level of human operations in space that is possible today.

The following excerpt taken from some NASA promotional material published before STS-1 highlights one of the claims made for the Space Shuttle before its first mission:

“Shuttle crews will be able to retrieve satellites from Earth orbit and repair and redeploy them or bring them back to Earth for refurbishment and reuse. The Shuttle can be used to carry out missions in which scientists and technicians conduct experiments in Earth orbit or service automated satellites already orbiting.” (1)

The key design characteristics of the Space Shuttle that made these claims possible 20 years ago included a heavy lift launch capability, a large payload bay, a robotic arm, the ability to carry a sizable crew, development of routine extra-vehicular activity (EVA) capability, and reusability of most key system components. Methodically, step-by-step during the last two decades, the ability of humans to work in space was expanded, increasingly more complex tasks were undertaken, and more and more of the resources of the crew and Shuttle were devoted to accomplishing the mission objectives.

Of historical interest perhaps are other claims made before STS-1. With respect to flight rate,

“The reusable Shuttle also has a short turnaround time. It can be refurbished and ready for another journey into space within weeks after landing.” (1)

With respect to mission operations,

“The Space Shuttle is basically an autonomous vehicle and ground support will be provided on an exception basis.” (1)

Neither the large flight rate nor the vehicle autonomy has been achieved. In fact, the greatest number of flights flown in one calendar year was eight and, although mission control has been streamlined since STS-1, particularly by using modern software applications and workstations, the issue of the appropriate division of labor between the ground and the flight crew is still debated today. Nevertheless, the vision of the Shuttle as a versatile platform for humans, hardware, and associated space operations was realized very much as advertised 20 years ago.

Orbital Flight Test Program

The first four Space Shuttle flights were designated the Orbital Flight Test Program (OFT) though the original planning in 1978 had specified six (2). OFT was specifically to demonstrate the basic performance of the orbiter vehicle as well as the external tank (ET) and solid rocket boosters (SRB's). Integrated together, these components are known as the Space Transportation System (STS). There were also performance requirements for several of the key subsystems and the need to demonstrate critical operational capabilities (2). One such critical objective was to demonstrate the novel approach to thermal protection. The Orbiter is covered with more than 30,000 insulating tiles and blankets that are designed to protect its aluminum structure from the heat of reentry. A critical design requirement specified that these tiles and blankets needed to be reusable. All previous manned spacecraft used nonreusable ablative material to protect the crew from the temperatures of reentry heating which reach several thousand °F.

STS-1 was launched on 12 April 1981. For the first time in NASA's history, a human-rated craft flew its initial flight with a crew on board. The risk of including the crew was viewed as necessary because their presence was critical to accomplishing the test-flight objectives.

Results from the first mission showed that the Orbiter could be launched, operated in orbit and landed safely, although Columbia glided farther in the landing phase than had been predicted. The Orbiter has no air breathing engines, so the commander makes the final approach and landing manually, flying "dead stick," without the option for a go around. Consequently, the glide ratio and handling qualities are crucial to routine, predictable landings. This was very much a test-flight program, which really had begun with the approach and landing tests in 1977. During the final drop test (the last of five) targeted to the hard surface runway at Edwards Air Force Base, the crew discovered that the landing phase control gains in Enterprise's software promoted a pilot-induced oscillation (PIO). This was apparent to the casual observer as the Enterprise oscillated dramatically during the final landing flare before eventually touched down safely. These gains were modified before STS-1, and the handling qualities were greatly improved. Better understanding of the aerodynamics of the Orbiter in the landing phase has enabled the crew to achieve pinpoint landings time after time.

Another important lesson learned came from the instrumentation pallet onboard Columbia on STS-1. It measured an acoustic environment at launch that was more severe than predicted and posed a threat to future payloads. For all subsequent launches, plastic "troughs" filled with water are installed on the mobile launch platform under the main engines to soften the impact of the acoustic shock of the main engine start. This has mitigated the potential damage to payloads that could result from engine start overpressure. Perhaps as significant as any demonstrated capability was the confirmation of the system of tiles and blankets. Although several tiles were lost from noncritical structure during Columbia's ascent, the thermal protection system worked as planned.

The second through fourth missions expanded the envelope and flight-tested other key systems for the first time. The most important flight test objective for STS-2 was to prove that the Shuttle was truly reusable. However,

another important step was the first “unloaded” use of the robot arm, as it was successfully flown without an attached payload. Incidentally, STS-2 was the last flight of a white external tank. All subsequent ET’s were left unpainted as a performance enhancement that saved approximately 600 pounds to orbit. STS-3 followed in early 1982 and demonstrated that the robot arm can handle a small payload that was unberthed and then reberthed in the cargo bay. STS-3 also expanded the allowable in-orbit thermal environment for the Shuttle.

The successful demonstration of the Remote Manipulator System (RMS), or robot arm, was critical to planned operations in space. It would be one of the ways to deploy satellites from the cargo bay and the primary way to retrieve free-floating satellites for in-orbit servicing or return to Earth. The RMS is manufactured from lightweight composite material, has six degrees of freedom, and each joint is electrically driven. The end of the RMS is shaped like a canister and contains snare wires designed to close upon a specially designed probe known as a grapple fixture. The RMS is not designed to capture objects that are not affixed with a grapple fixture. On Earth, the RMS could not lift its own weight, but for applications in space, it is certified for deployment of payloads up to about 60,000 pounds and retrieval of payloads of approximately 30,000 pounds. It is “flown” from the aft flight deck by a Shuttle astronaut, as shown in Fig. 1. Two hand controllers are used, one for rotation and one for translating a point of resolution (POR). The operator can select this POR through orbiter software as the end of the RMS or a point designated within an attached payload. It could also be the center of mass, a grapple fixture, a target, or some other reference point that has operational advantage. The operator selects a particular POR for a specific operation, and the Shuttle computer deconvolves the operator’s hand controller inputs into the appropriate motion of each of the joints to produce the commanded motion of the POR. In this manner, the operator “flies” the point of resolution. Software also can be used to control the rate at which joints move and thus ensure that when massive payloads are being maneuvered or operations are



Figure 1. The author is seen “flying” Discovery’s robot arm maneuvering EVA astronauts to various work sites on the Hubble Space Telescope during STS-82. Television monitors provide a view of the EVA activities in the payload bay. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

taking place in the vicinity of structure, the rates can be limited. This is important to protect against joint runaways and the possibility of collision or structural overload of the RMS. Backup modes are available in the event of failures. Each joint can be driven individually through a switch, with or without software support, although the operations are far more complex for astronauts if they are responsible for determining the motion of each joint on an individual basis to achieve a desired payload trajectory. This is accounted for in the training each operator receives preflight. The RMS also comes equipped with television cameras on the wrist and elbow joints which are used in conjunction with other cameras mounted at the four corners of the payload bay. These views are critical to the operator because in many cases a direct view out the aft windows may be blocked, for example, during the unberthing of a large payload.

The flight of a classified DOD payload on STS-4 completed the official OFT program. The USAF was an important future customer during the development of Space Shuttle design requirements and the potential uses of humans in space. STS-4 demonstrated the Shuttle's ability to land on a concrete runway, a key capability in the hoped for rapid turnaround that depended on the Orbiter's ability to land routinely at the Shuttle Landing Facility at KSC. Columbia returned to Edwards Air Force Base runway 22 on 4 July 1982 with President Reagan in attendance.

The OFT program completed these four missions in a little more than a year and demonstrated that the Shuttle could be launched safely, could conduct in-orbit operations including experiments, and return to a precise piloted landing. All key subsystems necessary for sustained in-orbit operations had been successfully verified and it was shown that the robotic arm could operate with an attached payload. The various simulators and training facilities were also validated.

After the OFT series was completed, the Space Shuttle was declared operational, but there was still much to be demonstrated. The fifth flight was the first to carry a crew of more than two and was the first demonstration that the Shuttle could serve as an in-orbit launch platform with the added value of a crew present to perform the last minute checkout of the satellite and preserve the option to return it to Earth if it could not perform its mission. STS-5 also had the first scheduled Shuttle spacewalk in space suits specifically designed for activity in zero gravity in or around the Shuttle. However, due to a suit system failure, the spacewalk on STS-5 was scrubbed. The EVA was successfully accomplished on STS-6, which also was the first flight of the second Shuttle in the fleet, Challenger. STS-6 also carried the heaviest payload, which was the tracking and data relay satellite (TDRS) satellite on an inertial upper stage (IUS). The Shuttle was the launch pad for the first of this constellation of satellites that would ultimately provide nearly complete space-to-ground communications coverage for the Shuttle and also for other satellites, including the Hubble Space Telescope.

Successful demonstration of Shuttle-based EVA was another key achievement in the operational maturity of the Program. The space suits designed for the Shuttle were composed of a hard upper torso (HUT), which came in a few discrete sizes. The lower torso assembly (LTA) consists of the waist and lower torso and boots along with the associated attachment rings. The Shuttle space suit is not custom built for each astronaut. The crew member is fitted with an

appropriate HUT and LTA, which can be adjusted. Collectively, all components are known as the Extravehicular Mobility Unit (EMU). The EMU was designed exclusively for operations in or around the Shuttle, not for excursions on a planetary surface, like the Apollo suits. Reusability was key to the design of the EVA system. A suite of standard EVA tools was developed that were designed primarily for Shuttle contingencies.

The completion of STS-6 more realistically represented the end of the orbital flight-test program. The next series of missions during the next 2 years began to develop more sophistication in human operations in space.

Rendezvous, Proximity Operations, and Satellite Servicing

By mid-1983, important operational techniques had been demonstrated but significant additional steps remained on the path to in-orbit satellite servicing. STS-7 saw the first flight of what was to become the standard crew of five astronauts. It was also the first flight of the astronauts from the class of 1978, the first group specifically selected for the Space Shuttle Program. This group included the traditional test pilot cadre and also scientists, doctors, and engineers who would perform the new job of Mission Specialist and be primarily responsible for EVA's, robotic activities, payload operations, and research. This group also included the first women and minority astronauts. STS-7's crew included the first American female astronaut to fly in space. A primary mission objective was to demonstrate the Shuttle's ability to deploy a satellite using the robot arm and then subsequently to retrieve it and return it to Earth. This was a critical operational skill to demonstrate for future in-orbit maintenance tasks such as the Solar Maximum Mission repair flight and later the Hubble Space Telescope missions. It was accomplished flawlessly and proved that the Shuttle could be precisely controlled near another orbiting vehicle. As a by-product of the operation, for the first time, a camera aboard the SPAS satellite provided a "bird's-eye view" of the Shuttle in orbit. Had weather permitted, STS-7 would have been the first landing of the Shuttle on the concrete runway at the Kennedy Space Center.

STS-8 made the first night launch and landing. Before this flight, launch times were picked so that mission landings would be made in daylight to reduce the commander's workload and maximize safety margins. However, night launches and landings would be necessary to the future of the Program for rendezvous with orbiting facilities such as the Hubble Space Telescope or a space station when launch times would be dictated by orbital mechanics. It was necessary to demonstrate that night landings could be trained for and executed safely. STS-8 was also the first demonstration that the robot arm could successfully unberth, maneuver, and reberth a payload weighing almost 8000 lb. Future satellites or space station components would be more massive still. The Hubble Space Telescope weighs about 24,000 lb on Earth.

STS-9 was the first flight of Spacelab, a large pressurized research module housed in the Shuttle's payload bay. This was the first demonstration that the Space Shuttle could serve as an orbiting laboratory for a multitude of experiments in many disciplines that took advantage of the microgravity environment, another promise made in the mid-1970s. Spacelab-1 was also the initial

opportunity for flights by Payload Specialists, a third category of crew member selected specifically because of expertise in the experiment complement of the specific mission. For the first time, the cadre of persons eligible to fly in space included nonprofessionals. The original design of the Space Shuttle was intended to provide an environment where nonprofessionals could participate in space flight. At 10 days, Spacelab-1 was the longest Shuttle mission to date and set the stage for the progressively longer missions that were to come. It was also the first flight that depended significantly on using the TDRS satellite, launched on STS-6, for the high data rate required for the 72 onboard experiments.

STS 41-B, the first flight in 1984, represented a significant step in increased complexity of in-orbit operations. Key mission objectives were to launch two communications satellites, conduct the first flights of the manned maneuvering units (MMU) that would be necessary for the upcoming servicing mission to the Solar Maximum Mission Satellite (SMM), conduct the first in-orbit rendezvous, and attempt once again the first landing at the Kennedy Space Center. The mission accomplished all objectives; however, the satellites that were successfully deployed from the Shuttle failed to achieve their final orbits due to similar failures in the nozzles of the Payload Assist Module (PAM) booster motor. This set the stage in the next year for the Shuttle to demonstrate its ability to retrieve satellites from low Earth orbit. The MMU had been test-flown inside the Skylab laboratory, but STS 41-B verified that a suited astronaut in space, untethered to the Shuttle, could fly the MMU. Using the MMU, the crew members positioned themselves with precision on attached fittings in the payload bay simulating the interfaces that the next crew would encounter on the SMM. The Challenger landed for the first time, as planned, at the Shuttle Landing Facility at KSC.

STS 41-C was a milestone in the Shuttle Program as NASA committed for the first time to attempt an in-orbit repair of a disabled satellite. Three years after its first flight the Shuttle realized some of its unique potential when Challenger served as an in-orbit platform to enable humans on site to fix the SMM. The Solar Maximum Mission Satellite was launched in 1980 but had suffered an electrical failure some months later. The Challenger would perform the first rendezvous from the ground to the proximity of the SMM, and an astronaut would fly the MMU from the Shuttle to the SMM and grapple a pin on the satellite with a trunnion pin attachment device (TPAD) attached to the MMU. Then he would fly the satellite back to the Challenger where it would be grappled by the robot arm and latched into a cradle in the payload bay. Unfortunately, the TPAD could not latch onto the satellite, and the repeated attempts to do so imparted a rotation of the SMM that made it impossible to retrieve. It was learned later that there had been a configuration change in the SMM that had not been generally known that affected the TPAD interface and did not allow the capture to take place. The initial attempt to retrieve SMM was abandoned, and ground controllers attempted to stabilize the spinning satellite.

There had been foresight on the part of the SMM team to equip the satellite with a grapple fixture that made it compatible with the RMS. Capture by the RMS would be the standard way that satellites would be retrieved in all subsequent in-orbit maintenance tasks, with some important exceptions to be discussed. Although not part of the original plan, the Shuttle robot arm was used to capture the free-flying satellite and berth it in the cradle, a technique that had been demonstrated

previously on STS-7. The successful retrieval was an example of NASA's maturity in real-time operations and dynamic planning. A new technique had to be developed and executed literally overnight because Challenger's propellant and SMM battery power were in short supply. This marked the beginning of an era in Shuttle operations where confidence in the vehicle and the crew allowed committing more of the margin to accomplishing the mission. Previously the priority was to preserve as much margin as possible, even at the potential expense of mission success, to assure crew and vehicle return. After the successful retrieval, the EVA crew repaired the satellite as planned. The satellite was released again by the robot arm and functioned flawlessly for the next several years (Fig. 2).

The next two flights continued to demonstrate that the Shuttle could serve as a launch pad in space. STS 41-D introduced the third Orbiter, Discovery, into the fleet. The mission successfully launched three communications satellites (the PAM nozzle problems experienced on STS 41-B had been fixed) and included the first flight of a commercial payload and commercial Payload Specialist. Only a few weeks later, STS 41-G launched with a crew of 7 for the first time and demonstrated a technique potentially important to in-orbit satellite servicing by having EVA astronauts transfer hydrazine fuel from a stowage volume to a simulated satellite fuel tank housed in the payload bay. This EVA included the first space walk by an American woman.

In November 1984, mission 61-A was launched to retrieve the two satellites that had been left in low Earth orbit by the failure of the PAMs on STS 41-B earlier in the year. Once again, the MMUs were used by an astronaut to fly to the stranded satellite where a probe affixed with a grapple fixture was inserted into the spent



Figure 2. Two Mission Specialist astronauts repair the “captured” Solar Maximum Mission satellite. The satellite is latched into a fixture at the aft end of Challenger’s payload bay. The astronauts use the RMS as a “cherry picker” for moving access to the satellite work site.

motor that was integral to the satellite (the PAM had been jettisoned earlier). Using the thrust from the MMU, the suited astronaut flew the satellite over to Discovery where the robot arm grasped the grapple fixture and maneuvered the satellite into the payload bay (Fig. 3). The procedure was repeated for the second satellite and both were returned to Earth in Discovery's payload bay (Fig. 4). They were each subsequently refurbished, sold, and launched successfully. However, this would turn out to be the last flight of the MMU's, and they were mothballed.

STS 51-D was launched on the fourth anniversary of STS-1 and was another flight in the series that launched communications satellites. One of these, SYNCOM IV-3, failed to activate after deployment from the Discovery. For the first time, the crew did an unscheduled rendezvous with the satellite and an unscheduled EVA in an attempt to activate a switch that was presumed to be the cause of the failure. Using materials found on board, including the plastic covers used on the checklists, the crew fabricated a pair of "fly swatters" which the EVA crew attached to the end of the RMS (Fig. 5). The "fly swatter" was used to snag a lever on the satellite that needed to be fully extended to initiate the satellite's sequencer. The operation was successful, but the root cause failure was elsewhere, and the satellite remained inactive. Four months later, Discovery returned on mission STS 51-I and repaired the disabled SYNCOM in orbit. The procedure required one of the EVA astronauts to grab the satellite while standing on a platform at the end of the robot arm. Appropriate fixtures were attached to the satellite, and it was handed off to the robot arm and repaired in the payload bay. SYNCOM was re-released, again manually, and the EVA crewman imparted



Figure 3. An astronaut wearing the MMU is seen approaching the Westar VI satellite that he will capture for return to Earth.

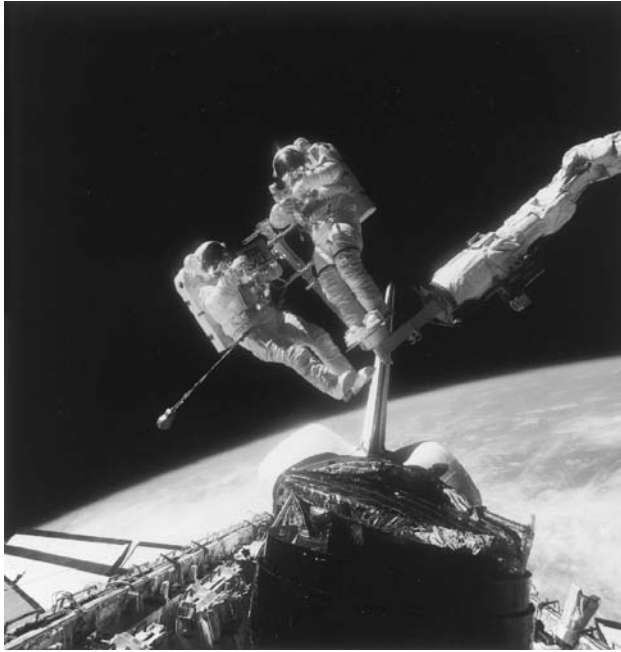


Figure 4. Two EVA astronauts hold a “for sale” sign after successfully retrieving two stranded communications satellites. The satellites were returned to Earth, refurbished, and relaunched.

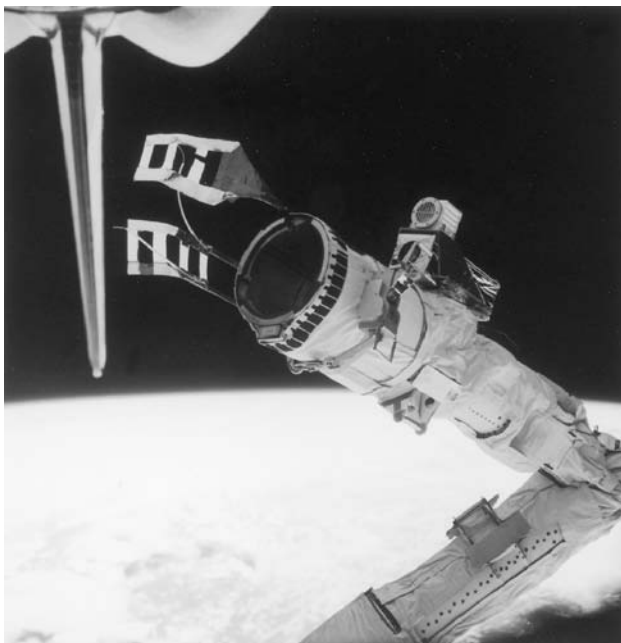


Figure 5. Two “fly swatter” attachments are seen strapped to the end effector of the RMS. The “fly swatters” were used in an attempt to activate a switch on an inert SYNCOM satellite deployed from Discovery in 1985.



Figure 6. An astronaut is seen standing in a foot restraint on the RMS while attempting to latch a handling fixture onto the INTELSAT- VI communications satellite.

a slow spin to stabilize the satellite before its solid rocket motor was fired. The satellite's subsequent in-orbit performance was normal.

In March 1990, an INTELSAT-VI satellite was left in a useless orbit by failure of its Titan rocket. Two years later, the Space Shuttle Endeavour joined the fleet and was launched on mission STS-49 to capture the satellite, outfit it with a new perigee kick motor, and launch it again. One space-suited crewman planned to ride on the end of the robot arm and attach a fixture to the satellite that would allow maneuvering the satellite into the payload bay where it would be subsequently mated with the booster stage. After a successful rendezvous with the INTELSAT-VI, the crew attempted the capture but without success because the latches on the fixture could not be made to capture the satellite (Fig. 6). Again, the ability of the ground and the crew to adapt in real time proved essential. A new plan was developed that involved using three EVA crew members on the same space walk for the first time in spaceflight history. Endeavour's commander re-rendezvoused with the spacecraft, stabilized the relative motion over the payload bay, and the three EVA astronauts simultaneously grabbed the free-floating satellite and maneuvered it by hand into the payload bay (Fig. 7). The subsequent rescue operations went as planned. The capabilities of the Shuttle and the presence of humans on site allowed completing the mission successfully. The INTELSAT-VI was deployed from Endeavour, and the motor burn successfully placed the satellite in an operationally useful orbit. The hardware and techniques to enable this repair mission had been developed very expeditiously and with the success of STS-49, the capability of in-orbit satellite rescue and maintenance reached a significant level of maturity.

In-Orbit Operations with the Great Observatories

At the time, new satellites were being designed to take advantage of the now demonstrated technique of in-orbit maintenance. Prime among these was the Hubble Space Telescope, the first of the Great Observatories. Originally envisioned



Figure 7. The first three-person EVA to capture the INTELSAT-VI satellite manually.

in the midtwentieth century and officially begun in the early 1970s, HST was to take full advantage of the capability of in-orbit repair and upgrade throughout its 15-year design lifetime. A comprehensive review of the history of the development of HST can be found in Reference 3. Unique to HST was the feature that many of its key components were designed from the outset to be replaced in orbit by space-suited astronauts. Failures were anticipated and provided for by this technique. It was also envisioned that as detector technology advanced throughout its lifetime, HST could continue to be a state-of-the art observatory by virtue of servicing missions that would install upgraded detectors. It would also be necessary throughout the lifetime of HST to raise the orbit continually. Normal atmospheric drag causes HST to descend slowly in time and because the telescope has no on-board propellant or thrusters, it has no way to combat orbital decay. However, the Shuttle's thrusters and propellant could be used to raise HST's orbit as necessary during the regularly scheduled servicing missions.

HST's launch was delayed for several years by the Challenger accident, but the time was well spent in planning for the initial deployment and subsequent in-orbit servicing. Ultimately, many components not originally designed for in-orbit change could be replaced. A set of tools specifically designed for working on the HST was fabricated and fit checked on the flight vehicle. Procedures were developed in the neutral buoyancy facilities that could be needed on either the deployment mission or the subsequent servicing missions. Ironically, the only major component that was impractical to replace other than the structure itself was the main mirror. The launch of the HST and the subsequent servicing is reviewed as the prime example of in-orbit maintenance, the value of humans in space, and the capabilities of the Space Shuttle fleet.

A discussion of the preparations for the initial deployment of the HST in 1990 is found in Reference 4. The capabilities of the Space Shuttle and the presence of a human crew were critical to the successful planning and execution of the mission. HST requirements dictated that the Shuttle should launch to the highest orbital altitude achievable. Rules were established that mandated a return of the HST if the insertion altitude were less than a minimum altitude that would allow for orbital decay of the HST with adequate margin for attitude control between the deployment mission and the first servicing mission that was scheduled approximately 3 years later. Although there was no concern that HST's orbit would decay to the point where the telescope would reenter, atmospheric drag was a threat to the rather feeble control authority of the reaction wheel assemblies that provide for HST's pointing. HST has no control jets. If HST could not point precisely enough to maintain sunlight on its arrays, the spacecraft would be at risk. To fit inside the payload bay, the solar arrays had to be folded up and stowed along the side of the telescope. The orbiter provided power to the HST while it was stowed. By necessity, the cable that provided the power was disconnected before deployment from the payload bay. So for the time interval until the solar arrays were deployed, the HST was only on its internal batteries. Should the solar arrays not deploy properly, HST would die. Therefore, the crew had to prepare to react immediately via EVA to deploy the arrays manually. There were additional, less time-critical procedures that required the presence of humans in space suits such as manually opening the aperture door. If all else failed, there were plans to bring HST back home for relaunch at a later time.

The HST was successfully launched on STS-31 in April 1990. The contemplated failure during the initial attempt to deploy the solar arrays actually occurred and necessitated that two of the crew don their space suits and enter the airlock. The ground team correctly analyzed the problem as a bad sensor, which was bypassed, and normal solar array deployment was completed without EVA intervention. All other mission events went as planned. It was approximately 6 weeks later that the world learned that HST had a flawed mirror.

The problem with the main mirror was a phenomenon known as spherical aberration (3). Had it not been for HST's design that allowed in-orbit servicing, the observatory's great potential would never have been realized, although even with the aberration, the telescope was doing important science. Based on an understanding of the true configuration of the main mirror, a solution could be engineered. The fix was to install a set of correcting optics within the telescope equipment bay that would precisely compensate for the mirror's true configuration. A new instrument was fabricated, called COSTAR (for corrective optics space telescope axial replacement), which had a series of correcting lenses and mirrors that would compensate for the HST mirror's incorrect curvature and would be inserted into the path of the light to the other instruments. COSTAR would be installed in the lower equipment bay of the HST in the location occupied by the high-speed photometer (HSP). The HSP would be removed and returned to Earth. A pointing stability problem also needed to be corrected. It was observed that the solar arrays would oscillate due to thermal effects as the HST passed from daylight to darkness (5), and compromised the stability of the HST. New solar arrays would be installed on the first servicing mission. To take

advantage of advances in detector technology, a new camera would also be installed to replace the original camera. This new camera would come complete with its own set of correcting optics.

When HST was launched, the first HST servicing mission, STS-61, was scheduled for about 3 years later. The value of the Space Shuttle and humans in space was probably never more evident than during the conduct of STS-61, the mission to save the Hubble Space Telescope. At the time, it was referred to as the most important mission NASA had flown since the Apollo 11 Moon landing. Some even asserted that the whole future of NASA was at stake.

To prepare for a mission such as this requires significant attention to detail in all aspects of mission planning and training. The crew simulates the maintenance tasks to be accomplished in space on Earth in a large pool. The training is highly realistic, although the astronaut is “neutrally buoyant” rather than truly weightless. The difference can be understood in the following way. The suits float, but the astronauts within do not. So depending on their orientation in the pool, they could float to the top of their helmets or firmly stand in the bottoms of their boots. The water’s natural viscosity also makes it easier to stop an object in motion than to start it, which is exactly opposite to the way things work in space. It can occasionally be awkward to use tools in the pool as well. Although it is possible for the astronaut trainee to be neutrally buoyant, it is more difficult to make tools that simulate weightlessness. The use of heavy tools in unusual orientations in the water can be particularly tiring for the hands and arms. Furthermore, the training suits are more flexible than the flightworthy suits. Suit stiffness is important because every time astronauts move, they have to displace air. In other words, the suit when inflated wants to resume its natural shape. It takes effort to do work in the suit and to fight constantly the 4.3 psi to which it is inflated. This is particularly true for hand-intensive tasks. Bending fingers and moving gloves can be very fatiguing. These training suits are not self-contained as are the real flight suits and so the breathing air and cooling water is provided to the astronaut through an umbilical. This umbilical must be managed and not allowed to interfere with the task.

Notwithstanding these differences, the excellent mock-ups that can be built make underwater training highly realistic. It is the current goal for astronauts to train 10 hours in the pool for every 1 hour of space walk scheduled. The pool comes equipped with its own robot arm allowing for integrated training between the space walking astronauts and the astronaut in control of the RMS. Most of the procedures involve using the robot arm as a work platform or “cherry picker” where the EVA astronaut stands on a platform mounted in the end effector and is maneuvered to and around the work site. This technique saves time, increases the work envelope, and reduces the EVA astronaut’s fatigue. It does, however, require a great deal of coordination.

Other training includes hands-on experience with the real hardware on the ground. This includes verifying the tool fit, the positions of handling fixtures, access, and other factors that affect the ease of performing the task in space. The HST was designed from the start so that all components designed to be changed in orbit would be “EVA friendly.” This means, among other things, large connectors that can be managed by a gloved hand, clear labeling, captive fasteners, and adequate places to position feet or hands. The HST is equipped with numerous

handrails that are painted yellow to designate them as allowable handholds unlike, for example, the magnetic torquers, which look a bit like handrails but are silver. The HST also has numerous receptacles on its surface for footholds called portable foot restraints, referred to as PFRs. These are essential for efficient EVA operations because they enable astronauts to use both hands for working rather than needing one hand to hold on to the telescope.

However, as previously discussed, it became desirable or necessary to change some components that had not been originally designed for in-orbit maintenance. In these cases, the astronauts can be faced with numerous small screws that must be managed or connectors that are hard to reach. It is a bit like trying to install the D-connector on the back of a personal computer while wearing oven mitts. Still, special tools and procedures can be designed. One such component is the solar array drive electronics (SADE) that was to be replaced on STS-61. This procedure required loosening numerous screws that were noncap-tive, but a tool was built that would capture the screws as they were removed as well as a carrier in which the screws could be stowed. The replacement SADE was designed with EVA-friendly connectors.

STS-61 was launched on 2 December 1993. A description of the mission from the perspective of one of the EVA crew members is found in Reference 6. During the subsequent 11 days, the crew performed a record five space walks and accomplished all planned tasks, including installation of the correcting optics that restored HST's eyesight along with NASA's reputation (Fig. 8). The careful

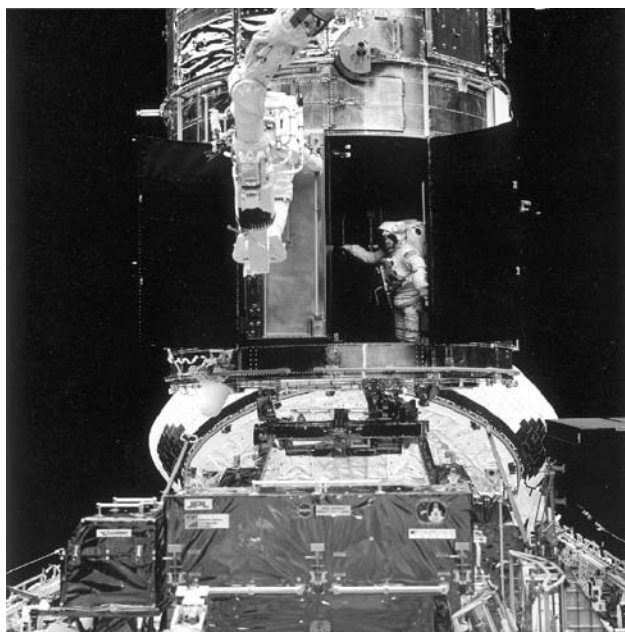


Figure 8. Mission Specialist astronauts prepare to install the COSTAR (corrective optics space telescope axial replacement) into the HST lower equipment bay during the first servicing mission. Note the astronaut holding the COSTAR while standing in a platform attached to the end of the RMS. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

preflight preparation paid off in the execution of the normal procedures as well when the crew was called upon to perform two unplanned tasks. On the first EVA, the bay doors that housed the replacement instrument would not close properly and the crew improvised by employing a strap to obtain additional leverage and successfully secure the door. Had the door not closed properly, the impact on the HST could have been catastrophic. During the second EVA, the crew replaced the original solar arrays. The preflight plan had the ground stow the solar arrays before replacing them. When one array bound during retraction, the crew was forced to jettison it manually. That exact operation had been the subject of the last simulation performed by the crew and ground controllers before launch.

In the end, the mission was a remarkable success and helped to define truly the importance of humans in space. In fact, the sensitivity of the new detector in the upgraded wide-field planetary camera more than compensated for the light loss due to the corrective optics. The original design specifications for HST's sensitivity and resolution were achieved.

If STS-61 was the Hubble Space Telescope repair mission, then the second HST servicing mission, STS-82, was the Hubble Space Telescope upgrade mission. Maintenance was only one dimension of the in-orbit servicing envisioned by HST's designers. It was also important to incorporate advances in detectors, and to that end, the second Hubble Servicing Mission was to install two new instruments to increase Hubble's spectral sensitivity from the ultraviolet to the near infrared (1100–20,000 Å). Hubble was launched with sensitivity in what is generally considered the "optical" part of the spectrum, about 1100–8000 Å (although the wide-field planetary camera had sensitivity to approximately 11,000 Å) (3). The two new instruments were the near-infrared camera and multi-object spectrometer (NICMOS) and the space telescope imaging spectrograph (STIS). These instruments are about the size of a refrigerator and are designed to slide into the equipment bay with very little clearance.

There was some concern before STS-82 that HST was working well enough to question the wisdom of doing the upgrade mission and the possibility of damage to the telescope. In the end, it was determined that the reward was worth the risk, and the mission was launched in February 1997. Many of the techniques used during this mission were identical to those employed on STS-61 servicing mission, and in five space walks, the crew successfully installed the new instruments and performed several other maintenance tasks. An insider's view of the second servicing mission is provided in Reference 7.

In another demonstration of the value of humans on site and NASA's ability to adapt in real time, it was observed on the approach to the HST that some of the thermal insulation on the side that preferentially faces the sun had become discolored, brittle in appearance, and loose in places. The thermal consequences were of concern as well as the possibility of debris resulting from lost pieces of insulation. The crew was given instructions how to fabricate replacement "blankets" and fasteners from materials flown onboard for other purposes, and one of the tasks during the fifth EVA was to install these temporary blankets over a few of the most distressed areas.

The original plan to maintain the HST's orbit envisioned using the Shuttle's two large orbital maneuvering engines. Due to load concerns, STS-61 reboosted

HST using a smaller set of maneuvering thrusters. However, due to concerns about the structural integrity of the solar arrays, even these thrusters would have imparted too much of a load when the solar arrays were extended, as they were during STS-82. It was necessary to have the crew execute a series of slightly uncoupled attitude maneuvers using the smallest 25-lb thrusters to raise the HST's orbit within the loading constraints of the arrays. Over the course of the STS-82 mission, the crew spent 82 minutes manually flying the orbiter to raise HST's orbit, although one reboost maneuver was also timed to serve as a collision avoidance maneuver between Discovery/HST and a spent upper stage. These vernier thruster reboost maneuvers have become the standard way the Shuttle is used to maintain the altitude of the International Space Station.

The results of the second servicing, or upgrade, mission were very successful. The new instruments performed as expected and produced astonishing new science. It was however, discovered later that there was a thermal short in the NICOMOS dewar. Solid nitrogen was used to cool the infrared detector, and the original design lifetime of the dewar system was 5 years. The thermal short was never cleared, and the nitrogen was gone in less than 3 years. The detector warmed, and the instrument became unusable. However, a cooling device was designed and flight-tested on a Shuttle mission in 1999. It is hoped that the cooler can be fitted to NICMOS on the HST servicing mission scheduled for early 2002 and thereby restore the infrared instrument's functionality.

Before the Hubble Space Telescope was launched in 1990, NASA had planned for the possibility of an "on demand" servicing mission caused, presumably by some component failure. Certain elements of the mission planning process would be completed in advance of the need, and the rest of the mission would be put together on an accelerated basis when called up. In 1999, the HST experienced a series of failures in its rate gyro units. The failures were due to a generic design deficiency and although the level of capability remaining was enough to keep HST safe, it was not sufficient to allow the telescope to do science. In response, NASA elected to accelerate a portion of the third servicing mission to launch as quickly as possible to restore the HST's science capability. This mission, known as servicing mission 3A, was devoted to the rate gyro replacement task, as well as replacing a failed transmitter, installing an upgraded fine guidance sensor, and replacing the original computer with one that was 20 times faster (8). Interestingly, the fine guidance sensor that was installed on the 3A servicing mission was the one removed by the second servicing mission crew on STS-82, which had been refurbished and outfitted with a correction for the main mirror's aberration.

In the future, NASA will continue to plan servicing missions at approximately 3-year intervals through the end of HST's original 15-year design lifetime. The next servicing mission, 3B, is scheduled for flight in early 2002. That mission will once again upgrade the solar arrays with arrays that are much more powerful and rigid. A new instrument, the Advanced Camera for Surveys will take the place of the wide-field planetary camera, and in a very ambitious maintenance task, the crew will attempt to change the power control unit (PCU). This procedure will be daring because of the requirement to power down the HST completely, something that has not been done in the more than its 11 years in orbit. As a matter of flight technique, the servicing tasks have always been

planned so that at the end of the crew's workday in-orbit, the telescope would be in a survivable configuration in the event that a Shuttle problem would necessitate early termination of the mission and, consequently, rapid release of the HST. In the case of the PCU change, it could be necessary to accept the risk of remaining "overnight" with the HST in a condition that should an orbiter emergency arise, the telescope would not survive. Again, however, this is symbolic of the level of confidence in the Space Shuttle's capability because it is more and more common to devote most of the flight margin to accomplishing the mission objectives, not hold them in reserve for vehicle and crew. As a matter of fact, a certain level of commitment is necessary to embark on any of these missions in the first place. A standard contingency that crews train to deal with is the loss of a propellant tank perhaps from collision with micrometeorites or orbital debris. In the case of impending loss of propellant, it is necessary to lower the Shuttle's orbit as soon as possible before the propellant leaks out. At the altitudes at which the Shuttle flies to service the HST, it would not be possible to return the telescope to a survivable configuration in time to respond to anything other than a small leak. In simplest terms, in the case of a propellant leak, it would come down to a choice of the telescope or the Shuttle.

The mission of the HST has been extended for another 5 years beyond the 15-year initial design lifetime. However, due to budget constraints, it is unlikely that new instruments will be built. It is hoped that servicing missions will continue as long as the telescope is doing world-class science and, at least, failures can be mitigated and orbital lifetime maintained. Perhaps at the end of its science life, the Shuttle will be used to return the HST to Earth where it might end up replacing its engineering replica, which is currently on display at the National Air and Space Museum.

Not all of the Great Observatories took full advantage of the presence of humans in space, but even when the satellite is not designed for in-orbit servicing, the crew can play an important role in facilitating the checkout and final deployment in low Earth orbit. The Gamma Ray Observatory (GRO), the second in the Great Observatories series, was launched aboard the Atlantis in 1991, almost exactly 1 year after the launch of the HST. Although not intended for in-orbit maintenance during its lifetime, the presence of the human crew was essential during deployment for the success of the mission. The high-gain antenna refused to respond to multiple ground commands to deploy, and the crew executed the second unscheduled EVA in Shuttle history to release the antenna boom manually. After its other systems checked out successfully, the GRO was released and burned on its own to a higher orbit where it successfully observed the most energetic phenomena in the cosmos, until it was intentionally deorbited in June 2000. The intentional destruction of the satellite was accomplished in response to failures in the telescope's attitude control system. Additional failures could have rendered the satellite uncontrollable and possibly a risk to inhabited areas if it were to reenter in that condition.

The third of NASA's Great Observatories was the Chandra X-ray Observatory (CXO, formerly, the Advanced X-ray Astronomy Facility, or AXAF). Originally to have been operating in low Earth orbit and serviceable by Shuttle astronauts, budget cuts in the early 1990s caused a redesign of the telescope's mission. Instead of making routine service calls as it does for the HST, the

Shuttle would serve as a launch platform and communications relay station. The CXO was destined to ride an IUS to a final orbit that would take it one-third of the way to the Moon. Before it was deployed from the Shuttle, its computers had to be loaded and the spacecraft had to be powered up. Had it not checked out properly, the CXO could have been returned to Earth. The Columbia carried Chandra to low Earth orbit in July 1999 where it was successfully activated and verified. The subsequent deployment and operation of the two stages of the IUS were flawless, and the Chandra has been sending back amazing images of some of the most energetic objects in the cosmos. Additional information on the mission to deploy the Chandra X-ray Observatory is available in Reference 9. A comprehensive review of the history of the development of the observatory and its operation has recently been published (10).

Assembling the International Space Station

The image of astronauts assembling a large research station in orbit has been in the minds of writers, scientists, and engineers for decades. Even in the mid-twentieth century it was clear that a large space station would necessarily require assembly by humans in-orbit, and the ability to operate there for long periods would offer the chance of optimizing the value of humans in space. For an insider's account of the history of NASA's development of the space station program, see Reference 11. The actual work to assemble the International Space Station (ISS) in orbit finally began in 1998 with the launch of the Russian Functional Cargo Block, "Zarya," and the U.S. Node, "Unity."

As ambitious as the Hubble Space Telescope servicing missions have been to plan and execute, the challenge of in-orbit assembly and activation of the International Space Station represents to many the greatest technological challenge ever undertaken. The STS-61 Hubble Servicing Mission was the most complex flight ever to be developed and executed, and before the first element launch in 1998, it was clear that NASA was facing the equivalent of four to six Hubble servicing missions each year during the assembly of the ISS. As adept as crews and flight controllers had become in the techniques of rendezvous, EVA, and robotics plus the associated training, there was no question that the future assembly tasks would require an even greater level of sophistication.

One way to depict the increase in difficulty that faced the team was what became known as the "EVA Wall" chart (Fig. 9). As the chart makes clear, the number of EVAs accomplished in the whole history of U.S. spaceflight is small compared to the number of EVAs necessary to assemble and maintain the ISS. Before STS-1, NASA had conducted 39 EVAs during the Gemini, Apollo, and Skylab programs. The first Shuttle-based EVA took place in 1982 during the STS-6 mission, and before the beginning of ISS assembly, there had been 44 EVAs. In addition, U.S. astronauts conducted 3 EVAs from Mir during the Phase 1 program. Through the completion of the Phase 2 portion of the ISS assembly (flight 7A), there had been 24 EVAs devoted to ISS assembly and maintenance. Through assembly complete in 2006, there will be an estimated 102 U.S. assembly and maintenance EVAs and an additional 66 Russian assembly and maintenance EVAs.

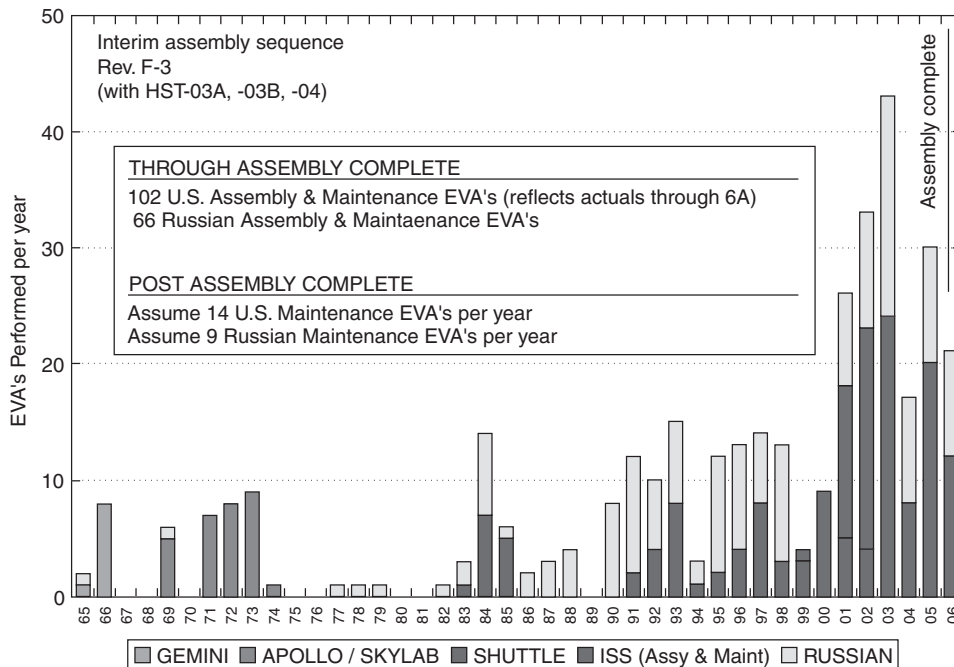


Figure 9. The chronology of space walks in the history of the U.S. space program and a forecast of the number of space walks that will be required to assemble and maintain the International Space Station. This is known as the “EVA Wall” chart. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The task of training for and accomplishing this series of EVAs is certainly difficult enough. Adding to the challenge is the fact that the hardware is being fabricated, checked out, and launched sequentially. This makes the task of verifying interfaces difficult or, in some cases, impossible. An unintended benefit of the schedule slip between the launch of the first ISS element and the beginning of assembly in earnest with the launch of the Z1 truss on STS-92 was the fact that flight hardware was continuing to be produced and shipped to KSC. This enabled a series of fit checks and interface verifications that provided increased confidence in the tools, techniques, and procedures before launch.

The sequence of flights from STS-101 to STS-104 included eight Shuttle missions, 18 EVAs, and spanned 14 months. In that time, NASA completed the assembly of a viable, self-sustaining in-orbit facility. With the installation of the P6 Solar Arrays, the “Destiny” Laboratory module, a robotic arm, and the “Quest” airlock, research on the ISS has begun, and Station crews can now conduct their own maintenance and assembly EVAs without having a Shuttle docked.

To date, the assembly sequence has gone remarkably well. Still, there have been some surprises and lessons learned. Once again, when problems arose, it was the humans in space and the knowledge and skills that they brought with them that remedied the situation. Some of what makes this possible is the

intimate knowledge of the hardware that the EVA crew has. Most have spent several years following the development and testing of the components that they are tasked to install. During the installation of the giant P6 solar arrays on STS-97, the tensioning mechanism failed to operate normally, and the crew used their knowledge of the design of the mechanism to make *in situ* repairs. Had repair not been accomplished, there was serious concern whether the arrays could take docking loads, and that would have compromised further assembly flights. Other in-orbit surprises included foot plates that interfered with the sockets into which they were to be installed, and an ammonia connector leak that occurred during the installation of the "Destiny" Lab. Given the complexity of the assembly tasks, these kinds of anomalies are considered minor and perhaps even expected.

There will no doubt be additional failures and process escapes with which future crews will have to deal. It will be their extensive training and experience with the flight hardware along with the experience of human operations in space that was developed during the Space Shuttle Program that gives the confidence that this staggering construction task can be successfully completed in the next few years.

The Future of Human Operations in Space

For more than 10 years, people have been discussing the type of vehicle that should replace the Shuttle. Many concepts have been examined, including Shuttle derivative vehicles with liquid boosters to replace the solid rocket motors. These new boosters would fly back autonomously and be reused. There are strong advocates of the single stage to orbit concept. Many of these proposed new vehicles make the same kind of claims made by Shuttle advocates in the 1970s, such as reusability, ease of maintenance, low cost to orbit, and robust flight rate. At the same time, NASA is developing the plans to continue operating the Shuttle fleet safely and effectively for another 20 years. A program of performance upgrades has been underway for several years that allow an even greater payload weight to orbit. Improvements in the crew displays and controls represent an important focus for upgrade investments in the next several years. Incremental improvements in key Shuttle capabilities such as robotics and EVA have been continually pursued throughout the last 20 years. The consistently low number of in-flight anomalies has demonstrated the reliability of the Orbiter itself. Systems reliability has been essential to accomplishing the more and more complicated missions. In the worst case, a systems problem could force the early termination of a mission and the loss of important mission objectives. Even a nuisance problem could tie up valuable crew time in troubleshooting and reconfiguration. The Shuttle's record of reliability is superior to that of any other launch vehicle. Even if no new human-rated vehicle is developed in the next decade, the investments in new technologies such as advanced propulsion, improved materials, and intelligent systems will offer the opportunity to improve the Shuttle fleet continuously.

The next level of sophistication in human operations in space will include developing the techniques to train for EVA and robotics operations in-orbit. ISS crews will not have the luxury of intensive training just before executing their tasks, as the Shuttle crews do today. More and more reliance will be placed on

skill-based, rather than procedure-based, training and training that can be conducted in flight. Increased sophistication in robotics will assist the human crew in accomplishing the more routine tasks. The increased sophistication of the ISS robot arm (known as the SSRMS) allows it to “walk” from one location to another across the surface of the “Destiny” Laboratory and maneuver much more massive objects than the Shuttle RMS. In the future, the SSRMS will translate on a mobile platform allowing the crew to have robotic reach throughout most of the entire growing ISS complex. Additional robotic aids will be developed that could eventually perform some of the more routine or repetitive tasks that currently require humans in space suits. ROBONAUT is a teleoperated, anthropomorphic robot that is being developed at the Johnson Space Center. ROBONAUT uses virtual reality techniques that allow for human operator control and is designed with two human-like arms and hands so that it can use existing EVA tools and interfaces to perform work outside the ISS complex.

It is certain that the incredible discoveries made by the Great Observatories during the last decade are due in large part to the role of human beings operating from the remarkable flying machine, the Space Shuttle. At one time, it was envisioned that satellites would be serviced and launched from a space station as is done now from the Shuttle. That vision may be years away still, but when humans begin again to explore the solar system in person, it may be that their craft will be assembled in low Earth orbit using the techniques envisioned by writers scientists and engineers in the early to midtwentieth century and developed in the last 20 years. When humans set foot on the surface of Mars, it will be the last two decades of the twentieth century that will be remembered as the time when man learned the critical operational skills that would make that historic voyage possible.

BIBLIOGRAPHY

1. NASA News Reference Summary. National Aeronautics and Space Administration, Washington, DC, 1981 pp. 1–5, 1–6, 9–5.
2. Shuttle Orbital Flight Test Generalized Test Logic and Progression Strategy. July 1978, *JSC* 14279.
3. Chaisson, E.J. *Hubble Wars*. Harper Collins, New York, 1994.
4. Hawley, S.A. *Sky and Telescope* 79: 373 (1990).
5. Feinberg, R.T. *Sky and Telescope* 86 (5): 16 (1993).
6. Hoffman, J.A. *Sky and Telescope* 86 (5): 23 (1993).
7. Hawley, S.A. *Sky and Telescope* 93 (2): 43 (1997).
8. Cohan, J.A. *Mercury* 30 (1): 24 (2000).
9. Hawley, S.A. *Sky and Telescope* 98 (2): 54–55 (1999).
10. Tucker, W., and K. Tucker. *Revealing the Universe*. Harvard University Press, Cambridge, 2001.
11. Mark, Hans, *Space Station A Personal Journey*. Duke University Press, Durham, NC, 1987.

STEVEN A. HAWLEY
NASA Johnson Space Center
Houston, Texas

IMMUNOLOGY AND INFECTION IN SPACE

This article on alterations of immune responses in space travel is intended to provide a review of the background and rationale for the hypothesis that immune function of humans on long-term space flight missions will be sufficiently altered that immune compromise occurs, thus rendering astronauts subject to opportunistic infections, particularly activation of latent viruses and the development of malignancies. The article reviews published information on the evidence of immune compromise in humans and animals flown in space and in human and animal earth-bound models of spaceflight. These alterations in immune function are discussed in the context of model secondary immunodeficiency states where external forces predispose humans to repetitive infection, debilitation, and cancer. An attempt will be made to quantitate the degree of immune impairment observed in space travelers and earth-bound equivalent model subjects and to predict what the astronauts on the Mars Mission might experience in terms of the occurrence of reactivated viral infections and the development of malignancy. Because this article includes many technical terms that may not be familiar to the reader, a glossary of immunological terms is included as Table 1 to facilitate reading.

The authors of this article are the odd couple. Dr. Sonnenfeld studied the effects of spaceflight on immune responses as a major focus of his entire research career, and Dr. Shearer came late to the field after studying congenital and acquired human immunodeficiency for three decades. Hopefully, this blend of master and novice in space-related immunology research will produce an article worthy of reader interest and remembrance.

Evidence of Altered Immunity in Human Space Travelers

After several years of experimentation, it has become clear, that spaceflight can have major effects on immune parameters. These effects are outlined here. The

Table 1. Glossary of Immunological Terms

Blastogenic responses	Stimulation of growth of T- and B-lymphocytes by plant proteins, e.g., phytohemagglutinin (PHA), concanavalin A (con A), pokeweed mitogen (PWM)
Antiorthostatic, hypodynamic, hypokinetic suspension	Rodent model for some effects of spaceflight on immune responses. Involves no weight-bearing on hindlimbs and head-down tilt that results in muscle and bone disuse and fluid shifts similar to those seen during spaceflight. Similar to chronic bedrest in humans
B-cells	Lymphocytes responsible for antibody production
Cellular immunity	Immunity based on direct effects of T-cells and macrophages on intracellular pathogens, including viruses, some bacteria, and tumor cells
CD3 + T-cells	Cells that express a general marker for T-cells
CD4 + T-cells	Cells that express a marker for helper (helping B-cells to make antibody) T-cell activity
CD8 + T-cells	Cells that express a marker for cytotoxic (destruction of target cells) T-cell activity
CD19 + /20 + B-cells	Mature B-cells
Colony-stimulating factors	Proteins produced by blood cells that stimulate growth of white blood cells (e.g., granulocytes, macrophages)
Complement	System of serum proteins that kill bacteria
Corticosterone	Stress hormone
Cytokines	The messengers of immune responses. Soluble substances that pass messages from cell to cell that regulate immune responses. Includes interferons (IFNs), interleukins (ILs), and tumor necrosis factors (TNFs)
Delayed type hypersensitivity (DTH) skin tests	<i>In vivo</i> measurement of ability of lymphocytes to react to antigens previously experienced
Epstein-Barr virus	Herpes virus that infects B-lymphocytes and remains dormant for years
Histocompatibility antigens	Cell surface markers on many tissues that are recognized by the immune system as the basis for graft rejection and also play a fundamental role in regulating immune responses
Humoral immunity	Immune responses regulated by antibody: primary targets are extracellular pathogens, including most bacteria
Immunoglobulins	Serum proteins (IgG, IgA, IgM) that contain specific antibodies; IgG compartment contains four subclasses (IgG, IgG ₂ , IgG ₃ , and IgG ₄)
Innate immunity	Nonspecific first-line defenses against foreign pathogens
Isohemagglutinin	Natural antibody to red blood cell type, for example, anti-A
Leukocyte	White blood cell
Leukocyte blastogenesis	Division of lymphocytes that requires cooperation with monocytes/macrophages that indicates that lymphocyte activity is functional
Lymphocyte	Type of white blood cell that is primarily responsible for initiating, regulating, and carrying out specific immune responses
Macrophage/Monocyte	Type of white blood cells that destroys foreign pathogens and activates lymphocytes to begin immune responses
Natural Killer Cells	Lymphocyte-like cells that carry out nonspecific destruction of tumor cells and virally infected cells
Oncogene	Gene that codes for tumor-related proteins
Phagocyte	Cells that eat bacteria (e.g., neutrophils, macrophages)
Phenotyping	Flow cytometry test that quantitates the amount of specific surface proteins on blood cells

Table 1. (Continued)

<i>Pneumocytis carinii</i>	Opportunistic parasite that causes pneumonia in immunocompromised humans
Polymerase chain reaction (PCR)	Test that measures nucleic acid content of plasma [e.g., HIV ribonucleic acid (RNA)]
Splenocytes	Lymphocytes in the spleen
Thymocytes	Lymphocytes in the thymus

mechanisms by which spaceflight induces changes in immune responses remain to be established. Possible factors that could cooperate to induce changes in immune responses during spaceflight include (1) exposure to microgravity, including possible direct effects and indirect effects on cells responsible for immune responses. This could include microgravity-induced changes in muscle and bone that have indirect effects on the immune system that result from alterations in such factors as levels of 1,25-dihydroxyvitamin-D₃. This vitamin interacts with macrophages, cells that play an important role in immune responses; (2) stress induced by spaceflight; (3) changes in circadian rhythm (sleep deprivation) induced by spaceflight; (4) changes in nutritional intake that occur during spaceflight; (5) radiation encountered during spaceflight; and (6) multiple other factors not yet identified. All of the body's physiological systems are closely integrated, so it is also possible that any change in one body system, such as the cardiovascular or musculoskeletal system, during and after spaceflight could have dramatic effects on the immune system.

Also not established is the biomedical significance of the changes in immune responses observed during and after spaceflight. Although it is clear that long-term changes in immune responses that occur on Earth (such as AIDS) can affect resistance to infection, the clinical significance of short-term changes in immune responses that occur after spaceflight remains to be established. However, as the era of very long-term spaceflight missions appears before us, the possibility that long-term spaceflight could induce long-term changes in immune responses, similar to those that cause difficulties on Earth, looms before us. This will require considerable study to define the potential problems as well as potential solutions.

Cellular Immunity. Numerous experiments during the last 20 years suggest that spaceflight affects immune responses. The bulk of observed effects has been on cell-mediated immunity. Although there is little doubt that spaceflight affects immune responses, the mechanisms involved and the biological/biomedical significance of these effects remain to be demonstrated.

The effects of spaceflight on *in vitro* immune responses have been most frequently demonstrated using the leukocyte blastogenesis model. Although it is clear that exposure of cultured human peripheral blood leukocytes to the spaceflight environment results in inhibited leukocyte blastogenesis, the mechanism(s) involved and the significance of these findings remain hotly debated (1–4). Additionally, difficulties in obtaining appropriate culture conditions in the spaceflight environment could also have affected results.

It has been shown that the ability of lymphocytes in cell cultures to divide after stimulation by nonspecific mitogens (leukocyte blastogenesis) dramatically

decreased, compared to ground-based and in-flight centrifuged (1 g) controls (1–4). These data have been used as evidence for a potential direct effect of microgravity on mammalian cells; however, multiple additional mechanisms could be involved (5,6). Additional mechanisms could involve alterations in the cell–cell interaction from changes in the fluid dynamics of cultures (i.e., indirect effects instead of direct effects of microgravity on the cells).

T-cell and macrophage functional activities that involved killing target cells were definitively altered after cell cultures had flown in space; in addition, enhanced production of interferon-alpha/beta (IFN- α/β) has been reported after culture of human peripheral blood leukocytes in space (7–13). The production of interferon-gamma, interleukin-1 (IL-1), interleukin-2 (IL-2), and tumor necrosis factor-alpha (TNF- α) from cultures of human peripheral blood leukocytes exposed to spaceflight was also altered, compared to controls (5,7,8,10,11,14,15).

It has been very difficult to obtain samples from human space travelers to evaluate the effects of spaceflight on immune responses. The relatively small number of astronauts and cosmonauts, as well as the multiple operational and experimental requirements on crews, has limited the number of studies possible. Nevertheless, individuals have been studied who have flown for short time periods, as well as on longer term flights of about a year's duration (15). Additionally, because of the constraints of space travel, most analyses have been carried out on samples obtained immediately after return of the crews to Earth. This has made isolating the effects of spaceflight on human immune responses difficult. It is possible that landing stresses, as well as readaptation to gravity, could have occurred before samples were obtainable. Only recently have limited experiments been carried out in flight.

Several alterations in immunological parameters have been reported after short-term spaceflight. Note, however, that although these alterations in immunity have been reported, the biological and biomedical significance of these changes have not been determined. It is not known whether immunity of the space traveler to infections or to tumors is compromised as a result of spaceflight-induced immune alterations. These studies have not been carried out because of the limited experimental capacity, to date. It is hoped that establishing a permanent International Space Station laboratory will allow additional exploration of this important question.

Alterations in immune parameters reported after spaceflight included the following: decreases in lymphocyte number, decreases in leukocyte blastogenesis, increases in leukocyte number, increases in numbers of B- and T-lymphocytes, decreases in monocytes, increases in helper T-lymphocytes, decreases in cytotoxic T-lymphocytes, and an increase in the ratio of CD4+/CD8+ lymphocytes (16–19). Variability has been reported from flight to flight (i.e., not all results have been consistently seen after every spaceflight). It is clear, however, that spaceflight, even as short as 1–2 weeks, does alter immune responses.

A few recent studies have not been compromised by analysis only after return to Earth. These have involved evaluating delayed hypersensitivity skin test (DTH) responses to recall antigens, such as purified protein derivative, when the individuals were tested in flight (20,21). These were also the first studies to show that spaceflight could affect the formation of immune responses in short-term (21) and long-term flight (20).

Humoral Immunity. Few studies have been conducted to determine the effects of space travel on human humoral immunity. This has been due more to the difficulty of carrying out any immunization studies on the limited astronaut/cosmonaut pool, than to the lack of interest.

To date, the only studies in this area have involved examining total serum immunoglobulin levels after the return of crews to Earth. In long-term studies, Russian scientists have reported increases in the level of serum immunoglobulins, particularly total serum IgA and IgM (14,15). Serum immunoglobulin levels were not altered during short-term U.S. Space Shuttle flights (19,22). The increased immunoglobulin levels observed after long-term spaceflight have been interpreted as indicating enhanced opportunity for the development of autoimmune diseases (14,15). The entire area of spaceflight effects on human humoral immunity is, obviously, understudied and requires additional attention.

Innate Immunity. The effects of spaceflight on innate human immunity have been studied in a limited fashion. Similar results have been reported for short-term and long-term flights, primarily by Russian investigators. Alterations that have been observed include decreases in the killing of target cells by natural killer (NK) cells and decreased IFN- α/β production (12,14,15).

Altered Immunity Produced in Ground-Based Human Models

Stress. Spaceflight travelers are in a potentially stressful environment, so stress exposure has been used as a ground-based model for studying the effects of spaceflight on immune responses. The most effective model for chronic stress exposure during spaceflight has been the academic stress model (23,24). In this model, first- and second-year medical students were exposed, as a result of their standard academic duties, to periods of stress (exams) and periods of relaxation (vacation). Because of the nature of the medical school curriculum, large numbers of students are exposed to the same stressors and relaxation periods. This makes it possible to follow the group longitudinally over time. Results using the academic stress model have shown alterations in immune responses similar to those observed after spaceflight, including decreased interferon production (23,24).

Exercise. It has been known for some time that prolonged moderate to extreme exercise produces some alterations in human immune responses (25–27). These changes include an increase in the innate components of immunity, such as neutrophil counts, but decreases in neutrophil bactericidal capacity, NK cell activity, and specific immune responses, such as IL-2 production, blast transformation, and cytotoxic T-cell numbers.

High Altitudes. High-altitude exposure has been used as a model for the effects of spaceflight on immune responses (17,28) because the conditions of air pressure, isolation, and stress are similar to those often found during spaceflight. The results of these studies showed depression in immune responses similar to those observed after spaceflight.

Bed Rest/Head-Down Tilt. Bedrest has been the most common model used to study the effects of spaceflight on immune responses (14,29,30). Bed-rest studies involve staying in a prone position for an extended period of time (days to

months) and result in muscle disuse and a lack of load-bearing of muscles required for support. Additionally, a 5–6° head-down tilt is often added to the model, which allows a shift of fluid to the head, similar to that during spaceflight (29). Although it is not possible to model microgravity in the field of Earth's gravity, the bed-rest model results in many physiological occurrences similar to those during spaceflight. Increases in IL-1 production and decreases in IL-2 production (11), changes in leukocyte subset distribution, and alterations in neutrophil and macrophage function similar to those that occur during space-flight were observed using this model (11).

Isolation/Antarctica. Isolation, both in placing individuals in isolation chambers or caves, as well as the Antarctic winter-over, has been used as a model to demonstrate the effects of spaceflight on immune responses (14,31–35). Isolation studies in caves and isolation chambers have yielded changes in immune responses similar to those observed during spaceflight, including decreased cytokine production and altered leukocyte subset distribution (14,32–35). The Antarctic winter-over study which was conducted during the 1970s did not show major effects on the frequency of viral respiratory infections, but the study was carried out using very limited facilities and the results were inconclusive (31).

Several studies that involved more than 240 subjects (in several winter experiments), were done by the Australian National Antarctic Research Expedition (ANARE) project. The studies demonstrated a statistically significant decrease in DTH skin tests to recall antigen on numerous occasions (36–39). These *in vivo* tests of immune suppression were correlated in a cohort of 17 subjects with their *in vitro* counterparts during the 1993 ANARE expedition to the Antarctic (40). This report confirmed the depression of cellular-mediated immunity by the DTH test and demonstrated a correlation with a 50% decreased phytohemagglutinin (PHA) proliferation test, altered production of inflammatory cytokines, induction of atypical mononucleosis, increased herpes virus shedding, and expansion of the EBV-infected B-lymphocyte population (40). These abnormal immune responses were seen in ANARE missions that involved total isolation and confinement in harsh winter conditions for 8 months. Interestingly, similar cohorts of subjects who wintered at the ANARE Station on Macquarie Island, which physically permits visitor exchanges during the winter months, did not demonstrate these lymphocyte changes (37,41).

Recent studies on humans isolated during the Antarctic winter have assessed regulatory cytokine responses (41b). The stress of this model system altered the proinflammatory and anti-inflammatory cytokine balance. IFN- γ (proinflammatory cytokine) levels in plasma were significantly elevated, while IL-10 and IL-1RA (anti-inflammatory cytokines) were significantly decreased in volunteers with reactivated virus infections (41b) (J.S. Butel, personal communication, June 13, 2002). Interpretation of these studies is consistent with T-cell activation that might be due to latent viruses, a hypothesis that holds importance for determining the risks of space travel.

Sleep Deprivation. Sleep deprivation is becoming another model of space travel because of the disturbances in sleep cycle experienced by astronauts in space travel. There are reports that immunoregulatory cytokines, such as IL-1, IFN- α , TNF- α , and interleukin-6 (IL-6) are important participants in maintaining

of the normal circadian sleep cycle. Human studies have demonstrated increased levels of leukocytes and NK cells in sleep deprivation (42–44).

Mounting evidence indicates that the inflammatory cytokines, interleukins (IL)-1 β , IL-6, and TNF- α are involved in wake–sleep regulation and that administering them to humans results in sleepiness and fatigue. IL-6 may regulate sleep by through stimulating the hypothalamic–pituitary–adrenal axis. TNF- α and IL-1 β regulate sleep by activating the NF κ B, a DNA-binding protein involved in transcription. Sleep deprivation and wakefulness result in building up the levels of IL-1 β , TNF- α , and activating of NF κ B in the cerebral cortex; the highest levels occur during daytime. TNF- α has two receptors, TNF- α R1 (p55) and TNF- α R2 (p75). Strong evidence suggests that sleep is regulated by the p55 receptor (45), not by the p75 receptor (46).

Humans subjected to 88 hours of total sleep deprivation exhibited significantly elevated plasma levels of the sleep-regulating cytokines TNF- α R1 and IL-6, as compared to subjects who were permitted periodic 2-hour naps (46b). These changes appeared to reflect alterations of the homeostatic drive for sleep, because they occurred only in the totally sleep-deprived subjects. Moreover, intermittent napping was suggested as a possible countermeasures approach to prolonged spaceflight.

Alterations in Immunity in Space-Flown Animals

A wide variety of spaceflight studies on immune responses have involved the use of experimental animals. These investigations have been of short duration, and the bulk of studies has used the rat (47–59). Experiments have involved both the Space Shuttle and the Russian Cosmos Biosatellite.

Lymphoid Tissue/Mass. The first studies indicating that immune responses might be altered by spaceflight were carried out by Russian investigators (49); they demonstrated an involution of the thymus in rats after flight. This finding was later confirmed in a Space Shuttle experiment (48,51).

T-Lymphocyte Reactivity. In rat experiments, as in human studies, the first experiments involved leukocyte blastogenesis using samples from rats flown in a Russian Cosmos mission. These studies showed a minimal effect of spaceflight on blastogenesis (60); compartmentalization of the effects of spaceflight on blastogenesis was later offered as an explanation for this finding. Additional experiments later showed that leukocytes from the lymph nodes of animals exhibited diminished blastogenic responses to concanavalin A (Con A), but not to PHA or B-cell mitogens (54,55). No changes were seen in the blastogenic responses of splenic cells from the same animals (55). When animals were dissected inflight, however, splenic cell blastogenesis decreased (52,61). This could indicate differences in the effects observed during and immediately after flight.

Functional Immunity. IFN- γ production was examined in an experiment on Space Shuttle mission SL-3 (50). The IFN- γ titer was greatly reduced in the culture supernatant fluids of cells obtained from flown animals, as compared to the titer from cells of ground-based controls. Interleukin-3 (IL-3) measurements made of the same culture supernatant fluids showed a change in levels. Later studies showed increases in the production of cytokines such as IL-3 (in both

thymocytes and splenocytes) and IL-6 (in thymocytes only) after spaceflight (53), indicating some compartment-specific effects of spaceflight on postflight cytokine production. Alterations in the production and/or secretion of other cytokines, such as IL-1 and IL-2, have varied. When changes are seen, normal levels return soon after landing (54,55,61). Cells from animals euthanized in flight showed decreased production of IL-1, IL-2, TNF- α , and TNF- β relative to cells from ground-based controls sacrificed under the same conditions as the animals that flew (52); this indicates that there were in-flight effects on cytokine production that might differ from those observed postflight.

Studies were also carried out on the effects of spaceflight on NK cell activity. Studies on cells taken immediately after landing from rats flown on the Russian Cosmos Biosatellite showed a decrease in the ability of splenic cells to kill YAC-1 cells, compared to the splenic cells of ground-based controls (56). The ability of NK cells from flown animals to kill K-562 cells was not affected (56). When rats were euthanized in-flight aboard the U.S. Space Shuttle mission (SLS-2) and cells were harvested at that time, contrasting results were seen (52). Splenic cells from the rats euthanized in space had decreased ability to kill both YAC-1 and K-562 target cells. When spleen cells from animals flown aboard SLS-2 but euthanized immediately upon return to Earth were tested, only the ability to kill YAC-1 targets was inhibited compared to controls, and there was no effect on the ability to kill K-562 targets (52). Again, in-flight results differed from postflight results.

Alterations due to spaceflight exposure were also found in experiments on colony-stimulating-factor (CSF) responsiveness and leukocyte subset distribution (61). The response of cells from flown rats to both granulocyte-macrophage colony-stimulating factor (GM-CSF) or macrophage colony-stimulating factor (M-CSF) was severely inhibited after spaceflight (57,58). Changes were seen in leukocyte subset distribution after spaceflight in rats on the Cosmos capsule and the U.S. Space Shuttle (i.e., increases in the level of CD4⁺ helper T-cells) (47,51,57,58).

Very limited studies using Rhesus monkeys flown in a Cosmos capsule have been conducted; results were similar to those seen in humans and rats, including inhibited blastogenesis responses and interferon production (61). Other flight experiments showed decreases in IL-1 levels, decreases in IL-2 receptor levels, and decreases in the response of bone marrow cells to GM-CSF (62).

Oncogene Activation/T-Cell. Possibly as a consequence of depressed immunity, rodents exposed to space travel, it was shown, have a marked increase in the oncogene-produced protein 53 in the dermal layer of the skin (63); this raises the question whether decreased immune surveillance increased the 53 oncogene activation and production of its protein. It is possible that spaceflight-induced decrements of CD4⁺ T-lymphocytes lead to viral activation of oncogenes or to an increase in space-radiation-induced changes in oncogene activation.

Altered Immunity Produced in Ground-Based Animal Model: Antiorthostatic Rodent Model

Antiorthostatic, hypokinetic, hypodynamic suspension is the leading ground-based rodent model, including similar physiological results to those seen during

spaceflight (64–66). The model is similar in principle to the chronic head-down tilt, bed-rest model for humans. It includes suspending rats or mice with no load bearing on the hind limbs and head-down tilt (usually 15–20°), creating a situation where there is bone and muscle disuse and a fluid shift to the head (67). The model has limitations, but it is the best available for simulating spaceflight for rodents.

Using the antiorthostatic suspension model, the following has been observed: (1) the blastogenic response of T-cells was inhibited in lymph nodes (68) and peripheral blood (69) but was enhanced in the spleen (68,70); (2) IFN- α/β production was inhibited in both mice (71,72) and rats (73); (3) increased IFN- γ production was observed using rat lymphocytes (74); (4) exposure to M-CSF of bone marrow cells from suspended rats to resulted in a diminished number of macrophage colonies, (68,75); (5) superoxide production was depressed in antiorthostatically suspended mice (76); (6) NK cell function in mice was not affected by antiorthostatic suspension (68,75); (7) neutrophil function was inhibited after antiorthostatic suspension (76) but was unaffected in rats (77) and left an unresolved situation; (8) leukocyte subset distribution and class II histocompatibility molecule expression were not affected by antiorthostatic suspension (69,70,75); (9) IL-1, IL-2, prostaglandin E2, and TNF- α production were not affected by anti-orthostatic suspension (69,70); and (10) there was no correlation of corticosterone levels with alterations in immune responses (70).

It has also been shown that antiorthostatic suspension of mice affects resistance to infection (78). Female mice that normally resist encephalomyocarditis D virus became infected after 4 days of antiorthostatic suspension (78). The decreased resistance correlated with the decreased interferon production seen after suspension. In contrast to the diminished resistance to viral infection, the suspended mice had increased immunologic memory and resistance to infection from *Listeria monocytogenes* (72,79). More recent work with this model system has demonstrated an unusual susceptibility to *Klebsiella pneumonia* challenge, showing an inability to clear bacteria, increased mortality, and decreased time to death in experimental animals as compared to control animals (79b). These are the only data available that indicate a possible correlation between spaceflight modeling-induced alterations in immune responses and changes in resistance to infection.

Human Immunodeficiency Secondary to Effects of Space Travel

Examples of Secondary Immunodeficiency Pertinent to Space Travel

Ionizing Radiation. The harmful effects of ionizing radiation on immune responses have been well demonstrated by the production of infections and tumors in irradiated experimental animals (80). Specific T-cell responses are primarily affected, although nonspecific factors, such as phagocyte function, may also be altered. Antigen processing by macrophages is particularly impaired by low-dose radiation. The effect of irradiation on antibody responses is generally minimal.

Cell-mediated immunity is significantly impaired by radiation. Prophylactic craniospinal irradiation in children who had acute lymphocytic leukemia

has produced lymphopenia and impaired lymphocyte responses to phytohemagglutinin (PHA) for more than a year after treatment (81,82). T-cell lymphocytopenia and B-cell lymphocytosis persisted for more than a year in irradiated patients who had Hodgkin's disease, in addition to impaired *in vitro* lymphocyte responses to mitogens and antigens that persisted for 10 years in some patients (83). Children whose thymus was irradiated for "enlargement" in the past suffered from a high rate of autoimmune disease (vasculitis, sarcoidosis, thyroiditis, inflammatory bowel disease) and thymoma (84).

When patients who had intractable rheumatoid arthritis were given fractionated total lymphoid irradiation (2000 to 3000 rad), there were pronounced alterations in immunoregulatory T-cells (85,86). Irradiation produced generalized lymphocytopenia, especially among T-helper cells ($CD4^+$ T-cells) (Table 2). Added to a decrease in lymphocyte reactivity to mitogens and antigens was a fourfold reduction in the ability of pokeweed mitogen (PWM)-stimulated lymphocytes to secrete IgM and IgG (85). Changes in T-cell subsets depend on the amount of radiation as well as the target area (87).

Although the exact dosage of radiation exposure to astronauts on long voyages is unknown, there is general consensus that solar flares and their consequent radiation pose the greatest risk. It is also generally believed that there is no effective shield from these solar flares of radiation. Consequently, the information gathered on the effects of radiation on the immune systems of Earth inhabitants who had accidental and therapeutic radiation will be extremely important in attempting to assess possible risk factors for long-term space travelers.

Malnutrition

Protein-Calorie Malnutrition (PCM). Malnutrition results from a deficiency of nutrients that disrupts normal physiological function. Poor nutrition

Table 2. Analysis of Lymphocyte Subsets in Patients who had Rheumatoid Arthritis before and after Total Lymphoid Irradiation^a

Cell feature	Controls	Patients	
		Before therapy	After therapy
Absolute lymphocyte count (No./mm ³)	2038 ± 73 ^b	1621 ± 133	573 ± 84
Percent CD3 ⁺ T-cells (absolute No./mm ³)	75 ± 2.3 (1520 ± 49)	63 ± 4.6 (922 ± 110)	49 ± 5.2 (345 ± 61)
Percent CD8 ⁺ T-cells (absolute No./mm ³)	23 ± 2.9 (469 ± 59)	18 ± 2.6 (246 ± 30)	27 ± 3.6 (182 ± 35)
Percent CD4 ⁺ T-cells (absolute No./mm ³)	43 ± 3.4 (870 ± 69)	46 ± 4.4 (680 ± 85)	23 ± 3.5 (152 ± 31)
CD4 ⁺ :CD8 ⁺ T-cell ratio	1.72 ± 0.9	2.22 ± 0.1	0.76 ± 0.23

^aModified from Kotzin B. L., S. Strober, E. G. Engleman, A. Calin, R. T. Hoppe, G. S. Kansas, C. P. Terrell, and H. S. Kaplan, Treatment of intractable rheumatoidarthritis with total lymphoid irradiation. Reprinted with permission from the *New Engl J. Med.* 305: 969-976 (1981).

^bResults expressed as the means ± standard errors. The total T-cell (CD3⁺) compartment is made up principally of CD4⁺ (helper) T-cells and CD8⁺ (cytotoxic) T-cells. There is a 78% loss of CD4⁺ T-cells vs. a 26% loss of CD8⁺ T-cells.

adversely affects many aspects of immune function (88), and it is the most common cause of secondary immunodeficiency in the world. The deficiencies are usually multiple and involve varying degrees of calorie and protein deprivation.

Poor nutrition markedly increases susceptibility to infection and results in increased morbidity and mortality from many infectious diseases. Nearly 50% of malnourished children who require hospital admission are infected. Bacterial infections are frequent (89,90), particularly pneumonia. Although the specific etiology of pneumonia is often difficult to establish, tuberculous, *Pneumocystis carinii*, and staphylococcal pneumonias are not uncommon. Gram-negative infections of the blood and gastrointestinal and urinary tracts are common.

Measles causes a particularly high mortality in these patients; the rash may be absent, and giant cell pneumonia is common (91). Herpes infections are often severe (92), and the prevalence of hepatitis B antigenemia is high (93). Many children have fungal and parasitic infections, including *P. carinii* (94).

The dynamic interplay between malnutrition and infection results in a vicious circle. Infection increases the need for calories and protein and at the same time causes debilitation; both of these accentuate the nutritional deficiency, making the patient even more vulnerable to infection.

Nonspecific Immune Factors. Several factors, including anatomic barriers, phagocytosis, lysozyme, interferon, and hormones, are involved in nonspecific host defense, but in malnutrition few have been completely characterized. Many of these factors are adversely affected (95), and some, such as localized defects of the mucous membranes, may be critical in the pathogenesis of respiratory, gastrointestinal, and urinary tract infections.

Anderson and investigators (96), who studied 37 malnourished African-American and Hispanic inner city infants, detected decreased migration of neutrophils in response to bacterial chemotactic factor and decreased adherence of neutrophils to surfaces. Moreover, these abnormalities were corrected by restoring adequate nutrition.

Serum immunoglobulin concentrations are preserved or increased in infants who had protein-calorie malnutrition because immunoglobulin synthesis is increased until terminal collapse, possibly as a survival mechanism (97). Some functional antibody responses have been reportedly impaired, particularly to vaccines (98). The most apparent abnormality of the B-cell system in malnutrition is the high level of serum IgE. A likely explanation is that nutritional deprivation affects the immunoregulatory CD4⁺ T-cells and leads to overproduction of IgE (99).

Morphological changes of the thymus in malnutrition have been noted by Dourov (100). Thymic atrophy is common and usually involves the cortex before the medulla. Thymic atrophy is distinguished by a greater loss of cells than a decrease in cell size. Also, thymic tissue does not show the capacity to regenerate readily when proper diet is restored. Impairment of cell-mediated immunity is the most common immunologic abnormality in protein-calorie malnutrition (101). Lymphopenia occurs in about 25% of the children who die from malnutrition. Tuberculin tests may be negative, and DTH reactions to other antigens are also impaired (95). The absolute number of T-cells and IL-2 production are diminished (102), and imbalances in immunoregulatory subsets of T-cells have

been observed in malnutrition (103). Severely malnourished infants may have an impaired lymphocyte response to PHA (102).

Vitamin A Deficiency. Isolated vitamin A deficiency is rare, but it is often seen in populations at risk for PCM (104). Several reports have demonstrated an important therapeutic impact of supplemental vitamin A on the mortality of malnourished subjects (105). A significant reduction in all-cause mortality was noted for treatment of hospitalized patients and in community-based supplementation studies (106). Rahmathullah and colleagues (107) documented that the greatest reduction in mortality occurred in children younger than 3 years of age.

The relationship of vitamin A deficiency to measles has been investigated most extensively. Measles remains a devastating disease in populations at high risk for malnutrition. Measles induces a depression of serum vitamin A levels (108). Treating severe measles with vitamin A greatly improves morbidity and reduces mortality (109). In addition, Coutoudis and colleagues (110) found that vitamin A treatment of measles allows a more rapidly reversed infection-induced lymphopenia. They also noted an increase in IgG antibodies to the measles virus in treated patients.

The role of vitamin A may augment nonspecific immunity by maintaining the physical and biological integrity of epithelial tissue (111). Semba and associates (112) found disturbed lymphocyte subsets in patients who had clinical evidence of vitamin A deficiency: lower CD4:CD8 ratios and lower proportions of the powerful na T-cells. Vitamin A supplementation reversed these abnormalities. Vitamin A is important in B-cell lymphopoiesis, as demonstrated by Buck and associates (114). Decreased NK cell activity has been documented in certain animal models of vitamin A deficiency (114).

Vitamin A deficiency also decreases the specific antibody response (115). For example, Semba and colleagues (116) reported increased tetanus toxoid titers in children at risk for malnutrition who were given vitamin A supplementation 2 weeks before vaccination. Animal studies also support the concept that vitamin A deficiency impairs primary antibody response (117,118).

Relevance to Space Travel. New emphasis is being placed on the importance of nutrition in long-term space travel because of the dearth of knowledge about the role of macro- and micronutrients in the well-being of astronauts (119). Because of the numerous potential side effects of long-term space travel, including those of proper absorption, metabolism, and use of foodstuffs, minerals, and vitamins, an appropriate concern is the potential impact of poor nutrition upon astronauts' immune responses. From a huge body of literature that goes back 100 years, it is well known that poor nutrition, weak resistance, and chronic infection lead to debilitation and early death (120). Working from the hypothesis that long-term space travel might create protein-calorie malnutrition or vitamin deficiency, it is essential to consider the impact of these forms of malnutrition upon the immune responses of astronauts.

Infection: AIDS. In less than 20 years, the human immunodeficiency virus (HIV) infection has been the most eloquent textbook of immunology for the entire world. Through the illustration of immune havoc, the incidence of opportunistic infections, the delicate balancing of HIV viral load with the CD4⁺ T-cell level, and the immunoreconstitution produced by antiretroviral medication, it is

now possible to quantitate the progressive loss of immune function more precisely with the appearance of the terminal condition of AIDS. Sixteen thousand adults are infected with HIV each day, and the total incidence is 40 million patients who have AIDS in the year 2000 (121). The lessons in immunity and infection that have been taught by the study of the AIDS epidemic find immediate application to many areas of human activity and medicine. Space immunology is no exception.

Immunopathogenesis. After the HIV virion enters target macrophages and T-cells, there is a repetitive cycle of immune activation and partial clearing of the virus but a continuous reinfection of immune cells by replicating virus (122) (Fig. 1). Up to 1×10^{11} virions and 2×10^9 CD4⁺ T-cells turn over each day, indicating the extraordinary capability of this virus and the immune system to do battle with each other (123–125) (Fig. 2). Several components of the immune system are active in attempting to stop the spread of HIV; cytotoxic T-cells (CTL)

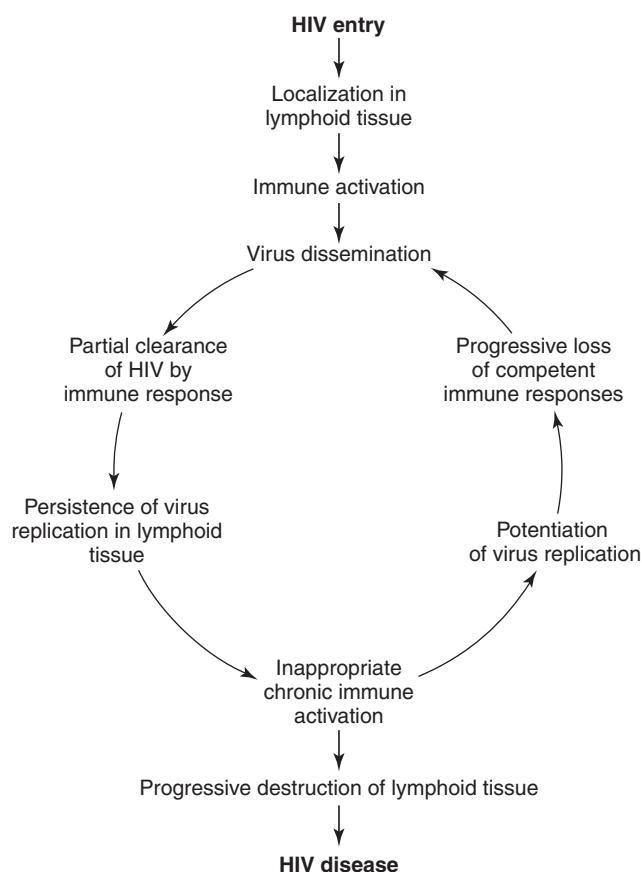


Figure 1. The relentless cycle of HIV infection. Taken from Pantaleo, G. and A.S. Fauci. New concepts in the immunopathogenesis of HIV infection. In W.E. Paul (ed.): *Annual Review of Immunology*, Vol. 13. Annual Reviews, Palo Alto, CA, 1995, pp. 487–512, with permission.

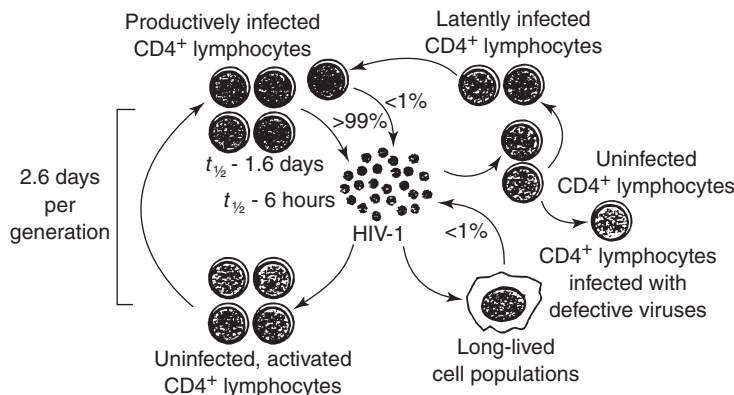


Figure 2. Dynamic interaction between HIV virions and CD4⁺ T-lymphocytes. Taken from Perelson A.S., A.U. Neumann, M. Markowitz, J.M. Leonard, and D.D. Ho, HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271: 1582–1586 (1996), with permission.

and neutralizing antibody (NAb) are the most prominent. Every part of the immune system is used by the abortive attempt to halt the spread of HIV infection. After 10–12 years, the components of immunity wear out, leading to severe immunosuppression, elevation of viral content in the blood, and the progression of HIV disease to death (126) (Fig. 3). The immune system actually contributes to its own defeat by producing cytokines that increase the spread of HIV infection (122,127). Inflammatory cytokines induce the activation of cells that promote the spread of HIV infection.

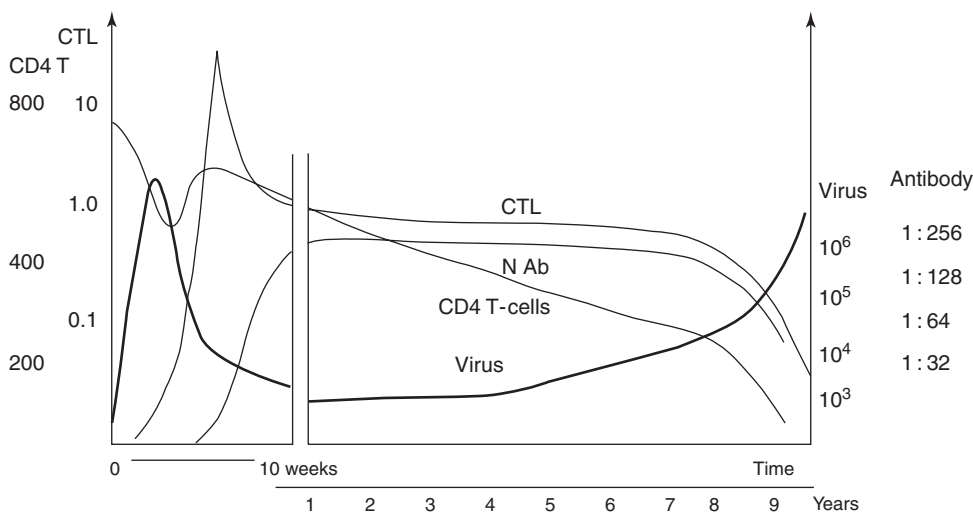


Figure 3. Time course of HIV infection in a hypothetical HIV-infected individual. Taken from McMichael A.J., and R.E. Phillips. Escape of human immunodeficiency virus for immune control. In W.E. Paul (ed.), *Annual Review of Immunology*, Vol. 15. Annual Reviews, Palo Alto, CA, 1997, pp. 271–296, with permission.

Viral Burden. By using the polymerase chain reaction (PCR) tests for HIV RNA, it is now possible to monitor more closely the natural history of HIV infection and to predict survival. The HIV RNA burden describes the onset of infection, the initial immune response, and the long-term prognosis of the disease (128,129). Soon after infection, most patients experience an acute rise in HIV, followed by a decline within 3 weeks due to anti-HIV immune responses and adjustment to a viral set point that is predictive of outcome (130). A high viral set point predicts the early onset of AIDS and death, whereas a low viral set point predicts relatively asymptomatic disease.

Use of Surrogate Markers to Determine HIV Disease Progression. Using the AIDS model, it has been possible to use surrogate markers to assess disease progression. When combination antiretroviral agents for treating of HIV infection were introduced, a number of assays were used to determine whether surrogate markers could substitute for the previously employed clinical end points. Of more than a dozen possible surrogate markers studied, two survived as useful: (1) HIV RNA PCR measurements and (2) CD4⁺ T-cell determinations. In addition to serving as independent predictors of HIV disease progression and response to antiretroviral therapy, these two measurements helped define the extremely active consumption of CD4⁺ T-cells and the release of HIV virions each day (131,132).

To validate a surrogate marker for use instead of clinical end points, several conditions must be met. There must be a relationship between the surrogate marker and disease pathogenesis and the mechanism of action of drug therapy. Using the AIDS model, it is evident that the HIV RNA burden measures the degree of infection, the CD4⁺ measures the extent of immunosuppression, and both of these surrogate markers improve in value with antiretroviral therapy. Another condition for validating the use of surrogate markers in clinical trials is that the surrogate markers should correlate with clinical outcomes across several trials. Changes in these surrogate markers must also mean the same thing clinically in treated versus untreated patients. Other requirements are that the measurement of surrogate markers should exhibit precision, reproducibility, dynamic range, feasibility, and cost-effectiveness. In addition, measurements of surrogate markers should be applicable to a wide spectrum of affected individuals.

Dependence upon surrogate markers in clinical trials of efficacy demands that laboratory performance of surrogate marker measurements address several crucial issues: consensus methodology, intra-assay variability, specimen handling ability, intrasubject variability, methods of quality assurance/quality control, and interlaboratory variability (132). Performance criteria for acceptable laboratory measurements must be established, continuing quality assurance/quality control of laboratory measurements must be made, and known specimens, must be sent out periodically from a central laboratory.

Quantitative Impairment of Immunity Due to Space Travel. Based on the human models of immunosuppression in therapeutic radiation, malnutrition, and AIDS, it is clear that a system for detecting alterations in the immune responses of space travelers needs to be devised. Because of the limitations of testing equipment and laboratory reagents, plus the lack of time and expertise in performing complex laboratory analysis in space, it surrogate markers of

Table 3. Categories of Immunosuppression in HIV Infection^a

CD4 ⁺ T-Cell Count/ μ L of blood	Degree of immunosuppression
> 500/ μ L	None
200–499/ μ L	Moderate
< 200 μ L	Severe

^aData taken from the Centers for Disease Control revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults, *MMWR* 41 (RR17): 1–19 (1992).

immune failing must be established and validated. At the moment, the most critical measurement of specific immunity is the CD4⁺ T-cell count, which again has been illustrated by the AIDS model (Table 3). Although this classification scheme for the degree of immunosuppression was discovered for a specific viral infection, it can be applied to other conditions of immunosuppression, such as cancer chemotherapy. When CD4⁺ T-cell counts fall to less than 200 cells/ μ L of blood, the individual is subject to opportunistic infections, no matter what the source of immunosuppression. Therefore, the ability to use this surrogate marker in astronauts in long-term space travel might be invaluable in determining serious alterations in immune protective responses. The technology to assess this important surrogate marker of protective immunity needs to be incorporated into long-term spaceflights, such as the journey to Mars.

Future Research Needs

Risk Assessment of Loss of Immunity: Specific and Nonspecific.

Based on our present imperfect knowledge of the extent of immunological changes that will take place in astronauts on prolonged spaceflights (e.g., Mars Mission) and the impact of these immunological changes on the health of astronauts, we must begin to attempt to quantitate the changes that have health consequences. Although clinicians have known for a century that infectious diseases of many types suppress immune responses, it was not until the AIDS epidemic came along that a massive campaign was mounted to study carefully the quantitative destruction of immunity and correlate it with the spread of the infection (Fig. 3). Similar quantitative immunologic data and clinical outcomes need to be gathered, ideally on astronauts, but more practically on Earth-bound subjects exposed to conditions that, it is thought, mimic those in space. There are two goals to pursuing these studies: (1) understanding the immunopathogenesis of spaceflight alterations in immunity and (2) developing countermeasures to avoid clinical consequences of spaceflight-induced immunosuppression. To begin to understand the extent of the immunologic consequences of spaceflight, we need to use the investigative tools developed during the last 50 years for humans who have developed immunodeficiency. This section has been included in this article to make clear to the reader the massive task before us and the tools that can be used to attempt this task.

In Earth-bound humans who pursue routine lives, a well-described series of examinations, procedures, and laboratory evaluations can detect the risk of loss of immunity. Physicians who specialize in diseases of immune dysfunction have available a series of screening tests and more sophisticated laboratory tests, by which it is possible to diagnose primary (congenital) and secondary (acquired) immunodeficiency diseases. Assessment of innate immunity is made by examining the polymorphic nuclear leukocyte, monocyte/macrophage, and complement components of immune function, whereas assessment of specific immunity can be determined by the relative strength of antibody and T-lymphocyte function. Figure 4 illustrates an algorithm that can be used to assess immune function quantitatively. Although the workup of the human immune function can be completed within just a few weeks, assessing the immune function in astronauts in long-term voyages into space poses special challenges.

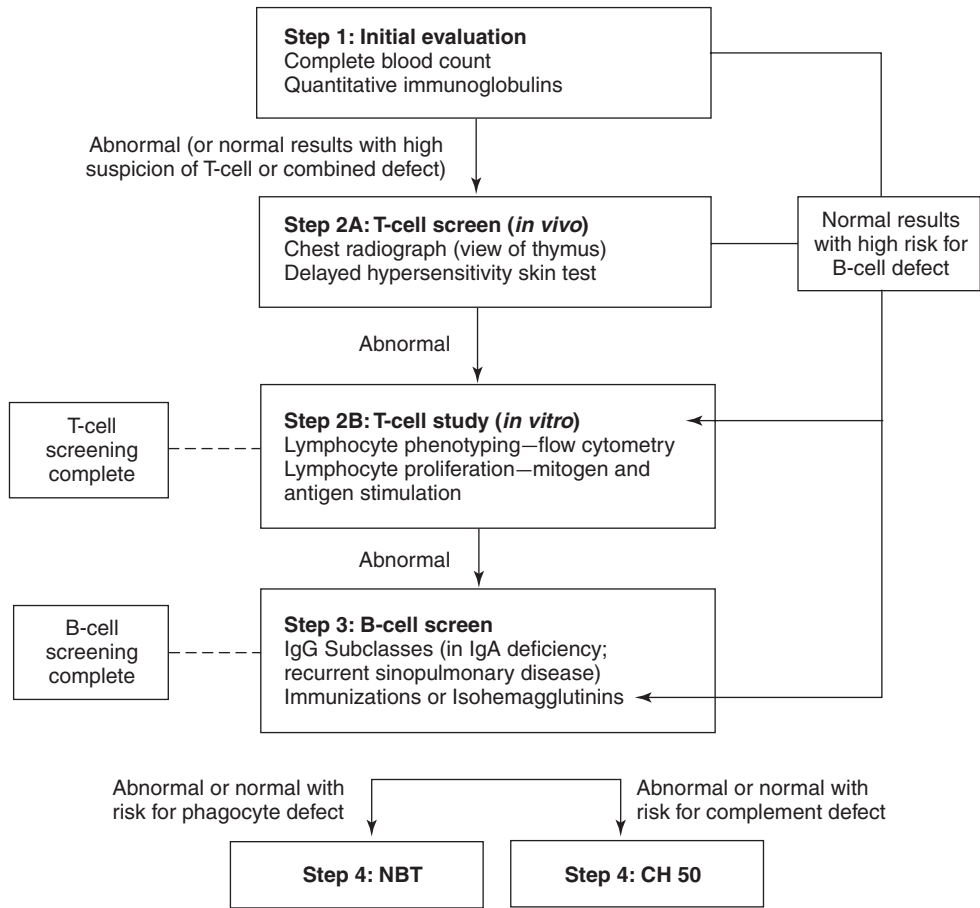


Figure 4. Stepwise approach to screening for immunodeficiency. Modified from Noroski, L.M., and W.T. Shearer, Short analytic review: screening for primary immunodeficiency in the clinical immunology laboratory. *Clin. Immunol. Immunopathology* 86: 237–245 (1998), with permission.

A "Stepwise Approach" to screening the immune system is proposed as a strategic method for evaluating each of the four major components of immunity.

Step 1: Initial Evaluation—Complete Blood Count and Quantitative Immunoglobulins. This is the first step in evaluating all suspected immune defects.

Complete Blood Count (CBC). A CBC is the first step in this approach; it offers meaningful information about several cell types (133–136). Analysis of absolute counts of leukocyte components compared to age-appropriate normative values is essential in the initial screening for immune defects.

Quantitative Immunoglobulins. When a B-cell is strongly suspected, levels of all immunoglobulins, IgG, IgM, IgA, and IgE, should be quantified (133,134,137).

When hypogammaglobulinemia is identified, it does not necessarily indicate antibody immunodeficiency (139). The diagnosis of antibody deficiency can be confirmed only when the patient fails to produce specific antibodies in response to antigenic challenge (138–140).

Step 2A: T-Cell Screen—Chest Radiograph and DTH Testing. Abnormal immunoglobulins or normal results in suspected T-cell or combined T- and B-cell defects are indications for specific T-cell assessment.

Chest Radiograph. Because immunodeficiency is most frequently associated with acute and chronic infections of the lungs, a chest X-ray is an excellent screening test for immunodeficiency.

DTH: In Vivo T-Cell Function. DTH is a reliable method of screening for cellular immune function; it is also cost-effective and can be performed in any setting (135,137). A positive DTH correlates with *in vitro* results of lymphocyte proliferative responses and cytokine production and is indicative of normal T-cell function (137).

The antigens commonly used are *Candida albicans*, tetanus toxoid, diphtheria toxoid, trichophyton, mumps virus, and tuberculin (137). A positive DTH is defined as a minimum of 5 millimeter diameter of indurated skin around the injection site and indicates antigen-specific T-cell infiltration; the test is maximum at 48 hours (139). This criterion applies to all of the antigens mentioned except tuberculin PPD for which there are specific thresholds for positive and negative results (140).

Adults who have been previously exposed to DTH antigens should respond to at least one of the antigens listed before (137,139). Skin test reactivity at 48 hours may indicate anergy, a cellular immune defect, or exposure to immunosuppressive agents, such as steroids (137). Subjects who do not respond should be retested using more concentrated forms of the antigens. Repetitive lack of response to DTH in the patient at risk for a T-cell defect warrants *in vitro* lymphocyte evaluation (137).

Step 2B: T-Cell Study—Lymphocyte Phenotyping and Proliferation Analyses. Lymphocyte functional assays are indicated when DTH is not normal. The determination of lymphocyte phenotyping by flow cytometry is necessary when the measurement of specific types of T-, B-, NK-, and phagocytic cells determines the diagnosis of a primary immune defect.

Lymphocyte Phenotyping: Flow Cytometry. Cell surface markers are identified by monoclonal antibody determination on the different T-, B-, myeloid, and

NK-cells of the immune system (141–146). By using monoclonal antibodies conjugated to distinguishing fluorochromes that emit light at different wavelengths, flow cytometry can be used to identify particular CD cell surface markers specific for immune cell types (143). The instrumentation, application, and interpretation of flow cytometry have been reviewed (137,141–146).

CD3 identifies panT-cells; CD4 marks MHC II restricted T-helper cells, and CD8 reflects MHC I-restricted T-cytotoxic cells. Surface markers of mature B-cells are CD19 and CD20. Absence of mature B-cells in the adult is associated with thymoma (139). In contrast, elevated levels of circulating B-cells, especially those that express CD5 (common T-cell antigen), is suspicious of chronic lymphocytic leukemia (138). Patients who have common variable immunodeficiency (CVID) have circulating B-cells, but they do not secrete immunoglobulin (137,138). Natural killer cells are commonly identified by CD16 marker (139,141–146).

Lymphocyte Proliferation: In Vitro T-Cell Function. Whole blood or density gradient separated lymphocyte samples are exposed to specific and nonspecific antigens in a sterile culture procedure. In normal subjects, the T-cells become activated, and new DNA synthesis occurs with lymphocyte mitosis in the presence of mitogens within 3 days or within 5 to 7 days in the presence of specific antigens (137). Transformation or proliferation is measured by the amount of radioactive thymidine incorporated into the DNA of the dividing lymphocytes. Mitogens such as Con-A, PHA, and PWM produce nonspecific proliferation of normal lymphocytes in culture. Commonly used T-dependent B-cell stimulants are PWM and fungi, whereas T-independent B-cell stimulants are EBV, *Staphylococcus aureus* (Cowan I strain), and lipopolysaccharide (137).

Proliferation of the lymphocytes indicates functional cellular immunity and correlates with *in vivo* DTH. Low or absent proliferative response usually indicates a cellular or combined antibody-cellular immunodeficiency.

Step 3: B-Cell Function—IgG Subclasses and Antigen-Specific Antibody Responses. This step is indicated in cases of suspected antibody deficiency.

IgG Subclasses. IgG Subclasses are obtained in those patients who have recurrent sinopulmonary infections and those who have IgA deficiency (137,147). IgG2 is usually low in those who cannot make antibodies to polysaccharide antigens. Comparisons must be made to standardized reference values for serum immunoglobulins and subclasses that reflect the patient's age group and the type of method used (134,137,138). Specific antigen lymphoproliferation should be confirmed in those who have with abnormal immunoglobulin or subclass levels.

Antigen-Specific Antibody Responses: Immunization and Isohemagglutinins. Specific humoral function is assessed by measuring antibodies to protein antigens after immunization with several antigens such as pertussis, tetanus or diphtheria toxoids, and polysaccharide antigens after immunization (137,138). If baseline antibody titers are low at initial testing, immunization should be performed and a blood sample obtained 3 weeks later (137). A fourfold or greater rise in a specific titer confirms a normal specific antibody response (138).

Antibody function can also be determined by isohemagglutinin assessment of naturally occurring IgM antibodies to type A and B red blood cells. Isohemagglutinins are naturally occurring IgM antibodies to type A and B red blood cells. A normal titer is at least 1:8 to A, B, or A and B blood groups (137). It is only

in AB blood type that this test is not meaningful because in this case the IgM antibodies would normally be absent.

Step 4: Phagocyte and Complement Evaluation

Phagocytes: Screen with NBT. Assessment of phagocyte function is reserved for those who have indolent infections, recurring severe staphylococcal wounds, or abnormal neutrophil counts. Evaluation for phagocyte defects should begin with testing the respiratory burst process by the nitroblue tetrazolium (NBT) test (137,139,148). Tests of oxidative metabolism include NBT, intracellular killing, and chemiluminescence. More sophisticated tests evaluate chemotaxis and measure phagocytic enzymes; flow cytometry identifies the lack of phagocytic surface markers (137,149).

Complement: Screen with CH50. Total hemolytic complement can rule out most complement defects and should be obtained in anyone who has severe, systemic infection with a bacterial pathogen (137) and tests the functional integrity of the classical complement pathway.

Summary of Four-Step Evaluation. Almost all suspected cases of immunodeficiency can be ruled out by appropriate interpretation of (1) complete blood count and immunoglobulin levels, (2) DTH, (3) isohemagglutinin measurement, and (4) complement lysis by CH50 (136).

Reactivation of Latent Viruses. Reactivation of latent viruses in the blood cells of astronauts exposed to long-term space travel poses a special threat because of the possibility of acquiring a debilitating chronic infection or malignancy. Viruses that produce these grave problems include EBV and HHV8. EBV is associated with human lymphomas and leiomyosarcomas, and HHV8 is associated with Kaposi sarcoma (150–154). Because long-term space voyages expose astronauts to stress, containment, isolation, microbial contamination, and radiation, it is not unreasonable to think that their immune systems would be weakened to the point that the normal immunosurveillance mechanism fails and latent viruses emerge from resident blood or tissue cells, leading to chronic inflammation and disease, and even to the appearance of frank malignancies.

Assessment of Innate Inflammatory Responses. To date, a detailed analysis of the effect of spaceflight on leukocyte recruitment has not been made. Considering that astronauts who travel on long-term space missions are likely to be exposed to a variety of inflammatory stimuli that will activate the inflammatory cascade, it becomes important to understand the potential effect of spaceflight on mechanisms of leukocyte recruitment and activation. It will be important to perform human and animal studies to examine directly the effect of simulated spaceflight conditions on inflammatory events, including leukocyte–endothelial cell interactions, adhesion molecules, and leukocyte activation.

Human Performance Trials and Immune Responses. Important experiments are needed in human performance studies to link the neuroendocrine–immune axis. Preliminary studies in sleep-deprived humans have demonstrated that normal circadian cytokine secretion cycles are altered (155,156). It is known that astronauts experience abnormal sleeping patterns during the stressful days of space travel. If the balance between TH1- and TH2-type cytokine secretion patterns were imbalanced by abnormal sleep patterns, it would be likely that a shift in the usual cytotoxic T-cell control could be replaced by the inflammatory T-helper cell response and would lead to a release of mediators such as TNF- α

and IL-10, which promote up-regulation of cell surface receptors for viruses and viral replication. The importance of precise circadian timing of the normal human sleep cycle and the impact of its disruption by sleepdeprivation have been confirmed (157).

In addition, chemokine receptor gene alteration and receptor expression need to be examined under these same conditions of sleep deprivation. It might be possible that once the cytotoxic T-cell response is lost, EBV and similar latent viruses find easy access to target cells of an increase in the surface expression of chemokine receptors. Stress-related immunomodulation and its implication for infectious diseases is receiving more attention based on publication of reports of longer recovery periods after infection in humans who had stressful conditions (158).

Synergy Research Projects in Radiative Biology, Virology, and Immunology. Radiative effects and containment of viral infection by the human immune system are likely to intersect. At some point, after significant radiation has been absorbed by T-cells, there is certain to be a threshold of weakness that permits viral infection to occur. Humans exposed to space radiation will, therefore, experience weakened T-cell immune responses that render them vulnerable to a number of common viral infections of humankind. Viruses carried within the human host cannot become activated, as is repeatedly shown in immunosuppressed humans. Collaborative research projects among space researchers in the fields of radiative biology, virology, and, immunology might help to anticipate severe infections and development of malignancy in human space travelers through collaborative research projects in animal subjects in Earth-bound space-equivalent models.

Countermeasures for Space-Related Immunodeficiency. The ultimate goal of all of these human studies is to determine whether astronauts in long-term space travel will experience alterations in normal immune function that immune resistance to infection and malignancy is impaired. If new information indicates a high degree of probability of immune impairment, countermeasures could be advocated for astronauts traveling to Mars. If specific antibody responses are likely to be impaired, special immunizations could be devised and intravenous immunoglobulin infusions could be supplied. If abnormal inflammatory cytokine production occurs in model studies, pharmacological agents could be given to inhibit the production of these mediators. Agents such as thalidomide and pentoxiphylline could be given to astronauts. If T-cell proliferation were likely to be significantly reduced, special immunizations with viral genome constructs in gene therapy could be given. These treatments are already being given to humans on Earth who have immunodeficiency, and these treatments could be prepared for delivery during long-term space travel.

Summary

The hypothesis for the proposals contained in this article has been that the effects of long-term space travel (isolation, containment, stress, microbial contamination, space radiation) would have a deleterious impact upon the human immune system. These harmful effects would predispose astronauts to chronic

infection and cancer. Our approach in developing this thesis has been to review the information available on the effects of space travel on the immune systems of astronauts and animals flown in space, to review the immune studies of space models on Earth, to quantitate the fall in immune resistance that produces opportunistic infections and premature malignancy, and to plan for developing of countermeasures to avoid these harmful effects.

ACKNOWLEDGMENTS

Dr. Shearer acknowledges the grant support of the National Biomedical Research Institute. Studies performed in Dr. Sonnenfeld's laboratory were funded, in part, by agreements and grants number NCC2-859 and NAG2-933 from the U.S. National Aeronautics and Space Administration.

BIBLIOGRAPHY

1. Cogoli, A., *J. Leukocyte Biol.* 54: 259–268 (1993).
2. Cogoli, A., and F.K. Gmünder. In S.E. Bonting (ed.), *Advances in Space Biology and Medicine*, Vol. 1. JAI Press, Greenwich, UK, 1991.
3. Cogoli, A., A. Tschopp, and P. Fuchs-Bislin. *Science* 255: 228–230 (1984).
4. Cogoli, A., M. Valuchi-Morf, M. Müller, and W. Breigleb. *Aviation Space Environ. Med.* 51: 29–34 (1980).
5. Bechter, H.A., M. Cogoli, O. Cogoli-Greuter, Müller, and E. Hunziger. *Biotechnol. Bioeng.* 40: 991–996 (1992).
6. Gmünder, F.K., M. Kiess, G. Sonnenfeld, J. Lee, and A. Cogoli. *Biol. Cell* 70: 33–38 (1990).
7. Armstrong, J.W., R.A. Gerren, and S.K. Chapes. *Exp. Cell Res.* 216: 160–168 (1995).
8. Chapes, S.K., D.R. Morrison, J.A. Guikema, M.L. Lewis, and B.S. Spooner. *J. Leukocyte Biol.* 52: 104–110 (1992).
9. Fleming, S.D., L.S. Edelman, and S.K. Chapes. *J. Leukocyte Biol.* 50: 69–76 (1991).
10. Limouse, M., S. Manie, I. Konstantinova, B. Ferrua, and L. Schaffar. *Exp. Cell Res.* 197: 82–86 (1991).
11. Schmitt, D.A., J.P. Hatton, C. Emond, D. Chaput, H. Paris, T. Levade, J.P. Cazenave, and L. Schaffar. *FASEB J.* 10: 1627–1634 (1996).
12. Talas, M., L. Batkai, I. Stoger, K. Nagy, L. Hiros, I. Konstantinova, M. Rykova, J. Mozogovava, O. Guseva, and V. Kozharinov. *Acta Astronaut.* 11: 379–386 (1984).
13. Woods, K.M., and S.K. Chapes. *Exp. Cell Res.* 211: 171–174 (1994).
14. Konstantinova, I.V., and B.B. Fuchs. *The Immune System in Space and Other Extreme Conditions*. Harwood Academic, Chur, Switzerland, 1991.
15. Konstantinova, I.V., M.V. Rykova, A.T. Lesnyak, and A.A. Antropova. *J. Leukocyte Biol.* 54: 189–201 (1993).
16. Meehan, R.T., L.S. Neale, E.T. Kraus, C.A. Stuart, M.L. Smith, N.M. Cintron, and C.F. Sams. *Immunology* 76: 491–497 (1992).
17. Meehan, R., P. Whitson, and C. Sams. *J. Leukocyte Biol.* 54: 236–244 (1993).
18. Taylor, G.R. *J. Leukocyte Biol.* 54: 202–208 (1993b).
19. Taylor, G.R., and J.R. Dardano. *J. Aviation Space Environ. Med.* 54: S55–S59 (1983).
20. Gmünder, F.K., I. Konstantinova, A. Cogoli, A. Lesnyak, W. Bogomolov, and A.W. Grachov. *Aviation Space Environ. Med.* 65: 419–423 (1994).
21. Taylor, G.R., and R.P. Janney. *J. Leukocyte Biol.* 51: 129–132 (1992).
22. Voss, E.W., Jr. *Science* 225: 214–215 (1984).

23. Glaser, R., J. Rice, C.E. Speicher, J.C. Stout, and J.K. Kiecolt-Glaser. *Behav. Neurosci.* 100: 675–678 (1986).
24. Kiecolt-Glaser, J.K., R. Glaser, E.C. Strain, J.C. Stout, K.L. Tarr, J.E. Holliday, and C.E. Speicher. *J. Behav. Med.* 1: 5–21 (1986).
25. Keast, D., K. Cameron, and A.R. Morton. *Sports Med.* 5: 248–267 (1988).
26. Nehlsen-Cannarella, S.L., D.C. Nieman, A.J. Balk-Lamberton, P.A. Markoff, D.B.W. Chritton, G. Gusewitch, and J.W. Lee. *Med. Sci. Sports Exercise* 23: 64–70 (1991).
27. Nieman, D.C., S.L. Nehlsen-Cannarella, P.A. Markoff, A.J. Balk-Lamberton, H. Yang, D.B.W. Chritton, J.W. Lee, and K. Arabatzis. *Int. J. Sports Med.* 11: 467–473 (1990).
28. Meehan, R.T., U. Duncan, L. Neale, G. Taylor, H. Muchmore, N. Scott, K. Ramsey, E. Smith, P. Rock, G. Goldblum, and C. Houston. *J. Clin. Immunol.* 8: 397–403 (1988).
29. Hargans, A.R., C.M. Tipton, P.D. Gollnick, S.J. Mubarak, B.J. Tucker, and W.H. Akeson. *J. Appl. Physiol.* 54: 1003–1009 (1983).
30. LeBlanc, A.D., V.S. Schneider, H.J. Evans, D.A. Englebretson, and J.M. Krebs. *J. Bone Miner. Res.* 5: 843–850 (1992).
31. Dick, E.C., A.D. Mandel, D.M. Warshaver, S.C. Conklin, and R.S. Jerde. *Antarctic J.* 12: 2–3 (1977).
32. Schmitt, D.A., C. Peres, G. Sonnenfeld, J. Tkackzuk, M. Arquier, G. Mauco, and E. Ohayon. *Brain Behav. Immunol.* 9: 70–77 (1995).
33. Schmitt, D.A., and L. Schaffar. *J. Leukocyte Biol.* 54: 209–213 (1993).
34. Sonnenfeld, G., J. Measel, M.R. Loken, J. Degioanni, S. Follini, A. Galvagno, and M. Montalbani. *J. Interferon Res.* 2: 75–81 (1992b).
35. Taylor, G.R., I.V. Konstantinova, G. Sonnenfeld, and R. Jennings. *Adv. Space Biol. Med.* 6: 1–32 (1997).
36. Williams, D.L., A. Climie, H.K. Muller, and D.J. Lugg. *J. Clin. Lab. Immunol.* 20: 43–49 (1986).
37. Muller, H.K., D.J. Lugg, and D. Quinn. *Immunol. Cell Biol.* 73: 316–320 (1995).
38. Muller, H.K., D.J. Lugg, H. Ursin, D. Quinn, and K. Donovan. *Pathology* 27: 186–190 (1995).
39. Pitson, G.A., D.J. Lugg, and H.K. Muller. *Arctic. Med. Res.* 55: 118–122 (1996).
40. Tingate, T.R., D.J. Lugg, and H.K. Muller. *Immunol. Cell Biol.* 75: 275–283 (1997).
- 41a. Lugg, D.E., *ANARE Medical Res.* in press.
- 41b. Shearer, W.T., B.-N. Lee, S.G. Cron, H.M. Rosenblatt, E.O. Smith, D.J. Lugg, P.M. Nickolls, R.M. Sharp, K. Rollings, and J.M. Reuben. *J. Allergy Clin. Immunol.* 109: 854–857 (2001).
42. Dinges, D.F., S.D. Douglas, L. Zaugg, D.E. Campbell, J.M. McMann, W.G. Whitehouse, E.C. Orne, S.C. Kapoor, E. Icaza, and M.T. Orne. *J. Clin. Invest.* 93: 1930–1939 (1994).
43. Dinges, D.F., S.D. Douglas, S. Hamarman, L. Zaugg, and S. Kapoor. *Adv Neuroimmunol.* 5: 97–110 (1995).
44. Dinges, D.F., and D.K. Chugh. Physiologic correlates of sleep deprivation. In J.M. Kinney, and H.N. Tucker (eds), *Physiology, Stress, and Malnutrition: Functional Correlates, Nutritional Intervention*. Lippincott-Raven, Philadelphia, 1997, pp. 1–27.
45. Fang, J., Y. Wang, and J.M. Krueger. *J. Neurosci.* 17: 5949–5955 (1997).
- 46a. Lancel, M., S. Mathias, T. Schiffelholz, C. Beh, and F. Holsboer. *Brain Res.* 770: 184–191 (1997).
- 46b. Shearer, W.T., J.M. Reuben, J.M. Mullington, N.J. Price, B.-N. Lee, E.O. Smith, M.P. Szuba, H.P.A. Van Dongen, and D.F. Dinges. *J. Allergy Clin. Immunol.* 107: 165–170 (2001).
47. Allebban, Z., A.T. Ichiki, L.A. Gibson, J.B. Jones, C.C. Congdon, and R.D. Lange. *J. Leukocyte Biol.* 55: 209–213 (1994).

48. Congdon, C.C., Z. Allebban, L.A. Gibson, A. Kaplansky, K.M. Strickland, T.L. Jago, D.L. Johnson, R.D. Lange, and A.T. Ichiki. *J. Appl. Physiol.* 81: 172–177 (1996).
49. Durnova, G.N., A.S. Kaplansky, and V.V. Portugalov. *Aviation Space Environ. Med.* 47: 588–591 (1976).
50. Gould, C.L., M. Lyte, J. Williams, A.D. Mandel, and G. Sonnenfeld. *Aviation Space Environ. Med.* 58: 983–986 (1987).
51. Ichiki, A.T., L.A. Gibson, T.L. Jago, K.M. Strickland, D.L. Johnson, R.D. Lange, and Z. Allebban. *J. Leukocyte Biol.* 60: 37–43 (1996).
52. Lesnyak, A., G. Sonnenfeld, L. Avery, I. Konstantinova, M. Rykova, D. Meshkov, and T. Orlova. *J. Appl. Physiol.* 81: 178–182 (1996).
53. Miller, E.S., D.A. Koebel, and G. Sonnenfeld. *J. Appl. Physiol.* 78: 810–813 (1995).
54. Nash, P.V., I.V. Konstantinova, B.B. Fuchs, A.L. Rakhmilevich, A.T. Lesnyak, and A.M. Mastro. *J. Appl. Physiol.* 73: 186S–190S (1992).
55. Nash, P.V., and A.M. Mastro. *Exp. Cell Res.* 202: 125–131 (1992).
56. Rykova, M.P., G. Sonnenfeld, A.T. Lesnyak, G.R. Taylor, D.O. Meshkov, A.D. Mandel, A.E. Medvedev, W.D. Berry, B.B. Fuchs, and I.V. Konstantinova. *J. Appl. Physiol.* 73 (Suppl.): 196S–200S (1992).
57. Sonnenfeld, G., A.D. Mandel, I.V. Konstantinova, W.D. Berry, G.R. Taylor, A.T. Lesnyak, B.B. Fuchs, and A.L. Rakhmilevich. *J. Appl. Physiol.* 73 (Suppl.): 191S–195S (1992a).
58. Sonnenfeld, G., A.D. Mandel, I.V. Konstantinova, G.R. Taylor, W.D. Berry, S.R. Wellhausen, A.T. Lesnyak, and B.B. Fuchs. *Aviation Space Environ. Med.* 61: 648–653 (1990).
59. Sonnenfeld, G., and E.S. Miller. *J. Leukocyte Biol.* 54: 253–258 (1993).
60. Mandel, A.D., and E. Balish. *Aviation Space Environ. Med.* 48: 1051–1057 (1977).
61. Lesnyak, A.T., G. Sonnenfeld, M.P. Rykova, D.O. Meshkov, A. Mastro, and I. Konstantinova. *J. Leukocyte Biol.* 54: 214–226 (1993).
62. Sonnenfeld, G., S. Davis, G.R. Taylor, A.D. Mandel, I.V. Konstantinova, A. Lesnyak, B.B. Fuchs, C. Peres, J. Tkackzuk, and D.A. Schmitt. *J. Interferon Cytokine Res.* 16: 409–415 (1996).
63. Ohnishi, T., N. Inoue, H. Matsumoto, T. Omatsu, Y. Ohira, and S. Nagaoka. *J. Appl. Physiol.* 81: 183–185 (1996).
64. Ilyan, E.A., and V.E. Novikovo. *Space Biol. Med.* 14: 128–129 (1980).
65. Morey, E., E. Sabelman, R. Turner, and D. Baylink. *Physiologist* 22: S23–S24 (1979).
66. Musacchia, X.J., D. Deavers, G. Meininger, and T. Davis, *J. Appl. Physiol.* 48: 470–476 (1980).
67. Taylor, G.R. *J. Leukocyte Biol.* 54: 179–188 (1993a).
68. Armstrong, J.W., K.A. Nelson, S.J. Simske, M.W. Luttgies, J.J. Iandolo, and S.K. Chapes. *J. Appl. Physiol.* 75: 2734–2739 (1993).
69. Nash, P.V., B.A. Bour, and A.M. Mastro. *Exp. Cell Res.* 195: 353–360 (1991).
70. Kopydlowski, K.M., D.S. Mcvey, K.M. Woods, J.J. Iandolo, and S.K. Chapes. *J. Leukocyte Biol.* 52: 202–208 (1992).
71. Rose, A., J.M. Steffen, X.J. Musacchia, A.D. Mandel, and G. Sonnenfeld. *Proc. Soc. Exp. Biol. Med.* 177: 253–256 (1984).
72. Miller, E.S., and G. Sonnenfeld. *J. Leukocyte Biol.* 55: 371–378 (1994).
73. Sonnenfeld, G., E.R. Morey-Holton, J.A. Williams, and A.D. Mandel. *J. Interferon Res.* 2: 467–470 (1982).
74. Berry, W.D., J.D. Murphy, B.A. Smith, G.R. Taylor, and G. Sonnenfeld. *J. Interferon Res.* 11: 243–249 (1991).
75. Chapes, S.K., A.M. Mastro, G. Sonnenfeld, and W.D. Berry. *J. Leukocyte Biol.* 54: 227–235 (1993).

76. Fleming, S.D., C.F. Rosenkrans, Jr., and S.K. Chapes. *Aviation Space Environ. Med.* 61: 327–332 (1990).
77. Miller, E.S., D.A. Koebel, S.A. Davis, J.B. Klein, K.R. Mcleish, D. Goldwater, and G. Sonnenfeld. *J. Appl. Physiol.* 76: 387–390 (1994).
78. Gould, C.L., and G. Sonnenfeld. *J. Biol. Regul. Homeostat Agents* 1: 33–36 (1987).
- 79a. Miller, E.S., and G. Sonnenfeld. *J. Leukocyte Biol.* 54: 578–583 (1993).
- 79b. Belay, T., H. Aviles, M. Vance, K. Fountin, and G. Sonnenfeld. *J. Allergy Clin. Immunol.*, in press.
80. United Nations Scientific Committee on the Effects of Atomic Radiation. *Ionizing Radiation: Levels and Effect*. United Nations, New York, 1972.
81. Campbell, A.C., P. Hersey, I.C. MacLennan, H.E. Kay, and M.C. Pike. *Br. Med. J.* 2: 385–388 (1973).
82. Petrini, B., J. Wasserman, S. Rotstein, and H. Blomgren. *J. Clin. Lab. Immunol.* 11: 159–160 (1983).
83. Kaplan, H.S., R.S. Hoppe, and S. Strober. Selective immunosuppressive effects of total lymphoid irradiation. In R.K. Chandra (ed.), *Primary and Secondary Immuno-deficiency Disorders*. Churchill Livingstone, New York, 1983, pp. 272–279.
84. Anderson, R.E., and N.L. Warner. *Adv. Immunol.* 24: 215–335 (1976).
85. Kotzin, B.L., S. Strober, E.G. Engleman, A. Calin, R.T. Hoppe, G.S. Kansas, C.P. Terrell, and H.S. Kaplan. *N. Engl. J. Med.* 305: 969–976 (1981).
86. Trentham, D.E., J.A. Belli, R.J. Anderson, J.A. Buckley, E.J. Goetzel, J.R. David, and K.F. Austen. *N. Engl. J. Med.* 305: 976–982 (1981).
87. Uh, S., S.M. Lee, H.T. Kim, Y. Chung, Y.H. Kim, C. Park, S.J. Huh, and H.B. Lee. *Chest* 105: 132–137 (1994).
88. Kuvibidila, S., L. Yu, D. Ode, and R.P. Warrier. The immune response in protein-energy malnutrition and single-nutrient deficiencies. In D.M. Klurfeld (ed.), *Nutrition and Immunology*. Plenum, New York, 1993, pp. 121–155.
89. Smythe, P.M., and J.A.H. Campbell. *South Afr Med. J.* 33: 777 (1959).
90. Phillips, I., and B. Wharton. *Br. Med. J.* 1: 407–409 (1968).
91. Smythe, P.M., G.G. Brereton Stiles, H.J. Grace, A. Mafoyane, M. Schonland, H.M. Coovadia, W.E. Loening, M.A. Parent, and G.H. Vos. *Lancet* 2: 939–943 (1971).
92. Templeton, A.C. *J. Clin. Pathol.* 23: 24–30 (1970).
93. Suskind, R.M., L.C. Olson, and R.E. Olson. *Pediatrics* 51: 525–530 (1973).
94. Hughes, W.T., R.A. Price, F. Sisko, W.S. Havron, A.G. Kafatos, M. Schonland, and P.M. Smythe. *Am. J. Dis. Child* 128: 44–52 (1974).
95. Neumann, C.G., C.J. Lawlor, Jr., E.R. Stiehm, M.E. Swenseid, C. Newton, J. Herbert, A.J. Ammann, and M. Jacob. *Am. J. Clin. Nutr.* 28: 89–104 (1975).
96. Anderson, D.C., G.S. Krishna, B.J. Hughes, M.L. Mace, A.A. Mintz, C.W. Smith, and B.L. Nichols. *J. Lab. Clin. Med.* 101: 881–895 (1983).
97. Chandra, R.K. Immunglobulins and antibody response in malnutrition—A review. In R. Suskind (ed.), *Malnutrition and the Immune Response*. Raven Press, New York, 1977, pp. 155–168.
98. Brown, R.E., and M. Katz. *Trop. Geogr. Med.* 18: 125–128 (1966).
99. Stiehm, E.R. *Fed. Proc.* 39: 3093–3097 (1980).
100. Dourov, N. *Curr. Top. Pathol.* 75: 127–150 (1986).
101. McMurray, D.N. *Prog. Food Nutr. Sci.* 8: 193–228 (1984).
102. Savendahl, L., and L.E. Underwood. *J. Clin. Endocrinol. Metab.* 82: 1177–1180 (1997).
103. Santos, J.I. *Infect. Dis. Clin. North Am.* 8: 243–267 (1994).
104. Rumore, M.M. *Clin. Pharm.* 12: 506–514 (1993).
105. Glasziou, P.P., and D.E. Mackerras. *Br. Med. J.* 306: 366–370 (1993).

106. Fawzi, W.W., T.C. Chalmers, M.G. Herrera, and F. Mosteller. *JAMA* 269: 898–903 (1993).
107. Rahmathullah, L., B.A. Underwood, R.D. Thulasiraj, R.C. Milton, K. Ramaswamy, R. Rahmathullah, and G. Babu. *N. Engl. J. Med.* 323: 929–935 (1990).
108. Arrieta, A.C., M. Zaleska, H.R. Stutman, and M.I. Marks. *J. Pediatr.* 121: 75–78 (1992).
109. Hussey, G.D., and M. Klein. *N. Engl. J. Med.* 323: 160–164 (1990).
110. Coutsoudis, A., P. Kiepiela, H.M. Coovadia, and M. Broughton. *Pediatr. Infect. Dis. J.* 11: 203–209 (1992).
111. Chandra, R.K. *Br. Med. J.* 297: 834–835 (1988).
112. Semba, R.D., Muhilal, B.J. Ward, D.E. Griffin, A.L. Scott, G. Natadisastra, K.P. West, Jr., and A. Sommer, *Lancet* 341: 5–8 (1993).
113. Buck, J., G. Ritter, L. Dannecker, V. Katta, S.L. Cohen, B.T. Chait, and U. Hammerling. *J. Exp. Med.* 171: 1613–1624 (1990).
114. Bowman, T.A., I.M. Goonewardene, A.M. Pasatiempo, A.C. Ross, and C.E. Taylor. *J. Nutr.* 120: 1264–1273 (1990).
115. Ross, A.C. *Proc. Soc. Exp. Biol. Med.* 200: 303–320 (1992).
116. Semba, R.D., Muhilal, A.L. Scott, G. Natadisastra, S. Wirasamita, L. Mele, E. Ridwan, K.P. West, Jr., and A. Sommer. *J. Nutr.* 122: 101–107 (1992).
117. Lavasa, S., L. Kumar, R.N. Chakravarti, and M. Kumar. *J. Exp. Biol.* 26: 431–435 (1988).
118. Pasatiempo, A.M., M. Kinoshita, C.E. Taylor, and A.C. Ross. *FASEB J.* 4: 2518–2527 (1990).
119. NSBRI Transition Activities. New Team Initiative: Nutrition, Physical Fitness, and Rapid Rehabilitation. April 11, 1999.
120. Sandberg, E.T., M.W. Kline, and W.T. Shearer. The secondary immunodeficiencies. In E.R. Stiehm (ed.), *Immunologic Disorders in Infants and Children*, 4th ed. W.B. Saunders, Philadelphia, 1996, pp. 553–602.
121. Shearer, W.T. HIV infection and AIDS. In W.T. Shearer (ed.), *Primary Care: Clinics in Office Practice. Allergy and Immunology*, Vol 25, No. 4. W.B Saunders, Philadelphia, 1998, pp. 759–774.
122. Pantaleo, G., and A.S. Fauci. New Concepts in the Immunopathogenesis of HIV Infection. In W.E. Paul (ed.), *Annual Review of Immunology*, Annual Reviews, Palo Alto, CA, 1995, pp. 487–512.
123. Ho, D.D., A.U. Neumann, A.S. Perelson, W. Chen, J.M. Leonard, and M. Markowitz. *Nature* 373: 123–126 (1995).
124. Perelson, A.S., A.U. Neumann, M. Markowitz, J.M. Leonard, and D.D. Ho. *Science* 271: 1582–1586 (1996).
125. Wei, X., S.K. Ghosh, M.E. Taylor, V.A. Johnson, E.A. Emini, P. Deutsch, J.D. Lifson, S. Bonhoeffer, M.A. Nowak, and B.H. Hahn. *Nature* 375: 117–122 (1995).
126. McMichael, A.J., and R.E. Phillips. Escape of human immunodeficiency virus for immune control. In W.E. Paul (ed.), *Annual Review of Immunology*. Annual Reviews, Palo Alto, CA, 1997, pp. 271–296.
127. Lee, B.N., J.G. Lu, M.W. Kline, M. Paul, M. Doyle, C. Kozinetz, W.T. Shearer, and J.M. Reuben. *Clin. Diag. Lab. Immunol.* 3: 493–499 (1996).
128. Mellors, J.W., C.R. Rinaldo, Jr., P. Gupta, R.M. White, J.A. Todd, and L.A. Kingsley. *Science* 272: 1167–1170 (1996).
129. Shearer, W.T., T.C. Quinn, P. LaRussa, J.F. Lew, L. Mofenson, S. Almy, K. Rich, E. Handelsman, C. Diaz, M. Pagano, V. Smeriglio, and L.A. Kalish. Women and Infants Transmission Study. *N. Engl. J. Med.* 336: 1337–1342 (1997).
130. Ho, D.D. *Science* 272: 1124–1125 (1996).

131. Mildvan, D., A. Landay, V. De Gruttola, S.G. Maschado, and J. Kagan. *Clin. Infect. Dis.* 24: 764–774 (1997).
132. Mofenson, L.M., D.R. Harris, K. Rich, W.A. Meyer III, J.S. Read, J. Moye, Jr., R.P. Nugent, J. Korelitz, J. Bethel, and S. Pahwa. *AIDS* 13: 31–39 (1999).
133. Shearer, W.T., R.H. Buckley, R.J.M. Engler, A.F. Finn, Jr., T.A. Fleisher, T.M. Freeman, H.G. Herrod III, A.I. Levison, M. Lopez, S.I. Rosenfeld, L.J. Rosenwasser. *Ann. Allergy Asthma Immunol.* 76: 282–294 (1996).
134. Shearer, W.T., M.E. Paul, C.W. Smith, and D.P. Huston. *Immunol. Allergy Clin. North America* 14 (2): 265–299 (1994).
135. Puck, J.M. *JAMA* 278: 1835–1841 (1997).
136. Paul, M.E., and W.T. Shearer. *Immunol. Allergy Clin. North America* 19 (2): 423–436 (1999).
137. Pacheco, S.E., and W.T. Shearer. *Pediatr. Clin. North America*. 41 (4): 623–655 (1994).
138. Lawton, A.R., and D.S. Hummell. ‘Primary antibody deficiencies’. In R.R. Rich, T.A. Fleisher, B.D. Schwartz, W.T. Shearer, and W. Strober (eds), *Clinical Immunology: Principles and Practice*. Mosby, St. Louis, 1996, pp. 621–636.
139. Paul, M.E., and W.T. Shearer. Approach to the evaluation of the immunodeficient patient. In R.R. Rich, T.A. Fleisher, B.D. Schwartz, W.T. Shearer, and W. Strober (eds), *Clinical Immunology Principles and Practice*. Mosby, St. Louis, 1996, pp. 609–620.
140. Committee on Infectious Diseases. American Academy of Pediatrics (AAP). Tuberculin Testing and Tuberculosis. In G. Peter (ed.), *1997 Red Book*, 24th ed. AAP, Elk Grove Village, Illinois, 1997.
141. Lanier, L.L., and A.L. Jackson. Monoclonal antibodies: Differentiation antigens expressed on leukocytes. In N.R. Rose, E.C. De Marcario, J.L. Fahey, H. Friedman, and G.M. Penn (eds), *Manual of Clinical Laboratory Immunology*, 4th ed. American Society for Microbiology, Washington DC, 1992, pp. 157–163.
142. Giorgi, J.V., A.M. Kesson, and C.C. Chou. Immunodeficiency and infectious diseases. In N.R. Rose, E.C. De Marcario, J.L. Fahey, H. Friedman, and G.M. Penn (eds), *Manual of Clinical Laboratory Immunology*, 4th ed. American Society for Microbiology, Washington DC, 1992, pp. 174–181.
143. Lewis, D.E., and W.J. Rickman. Methodology and quality control for flow cytometry. In N.R. Rose, E.C. De Macario, J.L. Fahey, H. Friedman, and G.M. Penn (eds), *Manual of Clinical Laboratory Immunology*, 4th ed. American Society for Microbiology, Washington DC, 1992, pp. 164–173.
144. IUIS/WHO Working Group. *Clin. Immunol. Immunopathology* 49: 478 (1988).
145. Comans-Bitter, W.M., R. de Groot, R. van den Beemd, H.J. Neijens, W.C. Hop, K. Groeneveld, H. Hooijkaas, and J.J.M. van Dongen. *J. Pediatr.* 130: 388–393 (1997).
146. Fleisher, T.A. *Immunol. Allergy Clin. North America* 14: 225–240 (1994).
147. Burrows, P.D., and M.D. Cooper. *Adv. Immunology* 65: 245–276 (1997).
148. Malech, H.L., and J.I. Gallin. *New Engl. J. Med.* 317: 687 (1987).
149. Abramson, S.L. Phagocyte deficiencies. In R.R. Rich, T.A. Fleisher, B.D. Schwartz, W.T. Shearer, and W. Strober (eds), *Clinical Immunology Principles and Practice*. Mosby, St. Louis, 1996, pp. 677–693.
150. Shearer, W.T., J. Ritz, M.J. Finegold, I.C. Guerra, H.M. Rosenblatt, D.E. Lewis, M.S. Pollack, L.H. Taber, C.V. Sumaya, and F.C. Grumet. *New Engl. J. Med.* 312: 1151–1159 (1985).
151. Paschall, V.L., L.A. Brown, E.C. Lawrence, R.A. Karol, E. Lotzova, B.S. Brown, and W.T. Shearer. *Pediatr. Res.* 18: 723–728 (1984).
152. Rosenblatt, H.M., D.E. Lewis, J. Sklar, M.L. Cleary, N. Parikh, N. Galili, J. Ritz, and W.T. Shearer. *Pediatr. Res.* 21: 331–337 (1987).

153. Stanley, S.K., and A.S. Fauci. Acquired immunodeficiency syndrome. In R.R. Rich, T.A. Fleisher, B.D. Schwartz, W.T. Shearer, and W. Strober (eds), *Principles of Clinical Immunology*. Mosby Year Book, St. Louis, 1996, pp. 707–738.
154. Kline, M.W., and W.T. Shearer. HIV infection and AIDS in children. In R.R. Rich, T.A. Fleisher, B.D. Schwartz, W.T. Shearer, and W. Strober (eds), *Principles of Clinical Immunology*. Mosby Year Book, St. Louis, 1996, pp. 739–750.
155. Dinges, D.F., W.T. Shearer, J.M. Reuben, J. Mullington, N. Price, M.P. Szuba, S. Kapoor, and H.P.A. Van Dongen. *World Fed. Sleep Res. Soc. Symp.* Dresden, Germany, October 5–9, 1999.
156. Shearer, W.T., J.M. Reuben, J. Mullington, N. Price, and D.F. Dinges. *Int. Forum Space Technol. Appl.* Albuquerque, NM, January 30–February 3, 2000.
157. Czeisler, C.A., J.F. Duffy, T.L. Shanahan, E.N. Brown, J.F. Mitchell, D.W. Rimmer, J.M. Ronda, E.J. Silva, J.S. Allan, J.S. Emens, D.J. Dijk, and R.E. Kronauer. *Science* 284: 2177–2181 (1999).
158. Glaser, R., B. Rabin, M. Chesney, S. Cohen, and B. Natelson. *JAMA* 281: 2268–2270 (1999).

WILLIAM T. SHEARER
Texas Children's Hospital
Baylor College of Medicine
Houston, Texas

GERALD SONNENFELD
Morehouse School of Medicine
Atlanta, Georgia

INTERNATIONAL SPACE STATION

The establishment of a long-term, low Earth orbiting outpost has long been a goal of space exploration advocates. A variety of orbiting space stations, envisioned by writers, engineers, and scientists for years, have been developed, launched, and deorbited. The Soviet Union and Russia have flown several Salyut stations and only recently terminated operations of the Mir Space Station (1). The United States flew three missions on the Skylab Space Station that was launched in 1975.

The long developmental history of the International Space Station (ISS) includes many lessons. The huge scale of the project offers many engineering, policy, and management lessons for our future endeavors in space. Despite the many challenges and the long development period, the project is now demonstrating great success. In November 2000, the first permanent crew docked with and entered the International Space Station (Fig. 1). Since then, the International team has continued to outfit and augment the ISS. At this time, the fifth expedition crew is well into their stay on board the ISS (Fig. 2). The results of many years of painstaking work are evident as the ISS flies overhead.



Figure 1. The first expedition crew onboard the ISS—Yuri Gidzeno, Bill Shepherd, and Sergei Krikalev (Photo NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

A Research Laboratory In-Orbit

From the time it was formulated, the ISS was meant to provide research capabilities not yet experienced in the microgravity environment of space. The space environment offers conditions that are not present on Earth or are not available at the same quality. These conditions include weightlessness, vacuum, space radiation, unobscured viewing of Earth and its atmosphere, and unobscured viewing of space (2). The ISS provides unique capabilities much greater than any past space research platforms. Power levels, “mass in orbit” assumptions, data capabilities, and the actual vibration-free, microgravity level drove the initial design concepts. All of these characteristics were codified as design requirements



Figure 2. The ISS configuration in-orbit as of June 2002 (Photo NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

for the ISS. In some cases, the requirements could not be met, and negotiations with science advisory groups have been conducted to reach the best solution.

Today, the ISS is an active research laboratory that has seven of its research racks in place. The ultimate complement for U.S. research racks is expected to be 27 racks of equipment that will cover many different disciplines of science, technology, and engineering research.

Our experience in using other space platforms has been limited either by a short-term mission or by scarce resources, such as power and the ability to have the scientists interact with the experiment. Our Shuttle experience has been limited to no more than 16 days in orbit. A world-class research facility must be able to conduct experimentation many more days under a number of conditions. The history of exploring this near Earth environment spans several decades; however, most of the actual experimentation has been done in just the past few years (3). Scientists have initiated studies in many areas of physics and material science. The lack of gravity-driven convection allows dramatic changes in many fields of research.

In biology, studies have been conducted at all levels: cellular, molecular, organisms, etc. Biological studies are required to identify the mechanisms for humans to survive in the space environment and also for the lessons that can be derived and applied to Earth-based problems. Many of the space effects on astronauts mimic disease on Earth. Understanding these mechanisms can provide important insights in to our problems here. In some cases such as protein crystallization and cell cultures, sample growth is larger and better formed in the space environment. The data from these samples can be more easily applied to current medical problems.

In engineering and technology areas, the ISS will provide a test bed for new technologies for spaceflight systems. The construction of the ISS itself is an engineering experiment that will impact the design of future spaceflight vehicles. The assembly and maintenance of this large facility will be critical for many future missions that will require assembly in space. The close proximity of the ISS to Earth also allows relatively rapid access to experimental data and samples (4).

In addition to the traditional basic research disciplines, a number of commercial enterprises are building equipment to use the ISS. Each of the ISS partners has research planning activities encompassing both traditional basic and applied research and the possible commercial research market. The commercial market is not yet mature. Although a number of studies have been done, lack of regularity in assured access will continue to limit commercial research in the near term. As the ISS is completed and operations develop into a well-understood regime, the commercial sector will be able to grow (5). Many reports and analyses of the commercial use of the ISS can be found through NASA's web site at <http://commercial.hq.nasa.gov>

ISS Development History

Early Studies. Many different concepts of U.S. space stations have been envisioned over the years; however, planning for the ISS began in earnest in the

early 1980s. The establishment of the space station task group under John Hodge set the groundwork for the selection of “reference designs” to be further analyzed. These designs, the power tower, the delta, the big tee, and the planar were used as reference configurations to evaluate against the use and operating scenarios that had been formulated. These studies led to designs that could be presented to President Reagan, the Administration staff, and other governments’ space agencies. The Space Station program was adopted as a new project in the fiscal year 1985 budget and was announced as an initiative in the President’s State of the Union address in 1984 (6). After integrating elements from the International Partners, a new configuration, the dual keel, was adopted (7).

At this point, NASA received funding for the Space Station from Congress, and planning began in the form of industry study contracts to begin the preliminary design. The Space Station design that formed the basis for these studies was the dual keel design. The Space Station would be constructed as a very large truss structure that had large upper and lower truss keels and a center core of laboratory and habitation modules. A mix of traditional solar arrays and two large solar dynamic power generators would provide electrical power. The overall configuration also contained a coorbiting satellite and a polar orbiting satellite.

Space station management was initiated with the Johnson Space Center (JSC) as the “lead center” of the project. JSC would manage the activities of four “work package” centers and perform the overall program integration and management. In fact, there would be no prime contractor; the NASA program office team at JSC would do the overall systems engineering and integration.

The four work package assignments were: Work Package 1—Marshall Space Flight Center (MSFC)—pressurized modules, life support and habitation systems. MSFC also included payload support equipment development and operations of payloads to include a payload operation and integration center (POIC). Work Package 2—JSC- truss elements and distributed systems such as the data management system; the communications and control systems; guidance, navigation, and control, and thermal control. JSC also included the development and management of new crew trainers, a new control center, and overall operations concept development. Work Package 3—Goddard Space Flight Center (GSFC)—development of the coorbiting and polar satellites, payload pointing equipment for the outside truss, a servicing bay on the truss, and the U.S. robotic capability the flight telerobotic servicer (FTS). Work Package 4—Lewis Research Center (LERC) now the Glenn Research Center (GRC)—electrical power generation and primary distribution.

In 1986, as Phase B study contracts were completed, the request for proposals for Phase C/D development and assembly contracts were released, and the contracts were awarded to industry in 1987. The Work Package 1 contract was awarded to a team led by Boeing. The Work Package 2 contract was awarded to McDonnell Douglas. The Work Package 3 contract was awarded to General Electric, and the Work Package 4 contract was awarded to the Rocketdyne division of Rockwell (now known as Boeing). In addition, the overall program office support contract was awarded to a Grumman-led team, and a contract for a Technical Management Information System (TMIS) was awarded to Boeing (8).

As a result of the Challenger accident in January 1986, NASA adopted recommendations from the Rogers Commission and a management review team

led by General Sam Philips that moved the program management function to the Washington, D.C. area for both the Space Shuttle and the Space Station (9). Overall program management responsibility for the Space Station was moved from the JSC to a new site in Reston, Virginia, 20 miles west of Washington, D.C. This move forced a difficult transition into the program during a challenging time. The industry proposals were being reviewed at all of the centers, and very few personnel made the transition from their home centers to the new Reston office. NASA spent months recruiting both inside and outside the agency for engineers, scientists, and administrative staff to man the new office. The move also highlighted a cultural issue within NASA. The NASA centers bristled under the management of an independent group. The Reston staff and the program manager were not given authority over the center staffs or the development contracts. The management concept was based on a weak centralized program office with only top-level oversight. Over the years, this would be a major criticism of this phase of the Space Station program (10).

During late 1986 and early 1987, NASA management was also concerned that the Space Station did not have a firm operating concept that would drive the preliminary design activities. A new team was established under Dr. Peter Lyman of the Jet Propulsion Laboratory (JPL) and Carl Shelley of the JSC to develop and document an integrated space station operating concept. The Space Station Operations Task Force (SSOTF) conducted reviews of existing NASA plans and interviewed academia, industry, and other government bodies to highlight the best operating concepts. The concepts included integrated mission planning, data system integration, manifesting, ground operations, and space operations. The results of the SSOTF were used to establish the overall operating architecture for the Space Station program (11).

The Reston program office was active from 1987–1993. During these years, the Space Station was given its first name, Freedom, and the initial agreements with our International Partners were signed. President Reagan's call to our allies to join the program had been answered by the Canadian Space Agency, the European Space Agency, and the Japanese Space Agency. In 1988, the Intergovernmental Agreement (IGA) and the individual Memoranda of Understanding (MOUs) were signed by all of the partners.

As the Freedom program began to require increased funding for the development contracts, the politics of the program became contentious. The U.S. Congress insisted on multiple reviews by the National Research Council and limited NASA's funding for Freedom several times during these years. The program had already encountered redesigns for technical considerations and now faced significant changes due to funding limitations. The upper and lower keels were separated into Phase 2 of the program and eventually dropped along with the servicing facility, the polar and coorbiting platforms, the solar dynamic power systems, a new extravehicular activity (EVA) suit, the U.S. robotic systems, and the payload pointing systems. Ultimately, in 1991, all of the content of Work Package 3 had been deleted, and the General Electric contract was terminated.

In addition, Congress imposed technical constraints. The amount of power available for research was fixed. The laboratory module must fly before the habitation module. The station would have a man-tended stage prior to permanent manning. These and several other Congressional constraints drove the design

and the deployment or assembly sequence of the Space Station Freedom program. Technically, significant progress was being made as the program proceeded through Preliminary Requirements Review (PRR), the Integrated System Preliminary Design Review (ISPDR) and finally a man-tended stage Critical Design Review (MTCDR). These milestones supported the detailed design and development of much of the hardware that is flying now as the ISS.

In late 1992 a large increase in the Freedom cost estimate was projected. The increase raised concerns in both the Congress and the new Administration about the management of the Freedom Program. Despite the volume of detailed design work that had been completed, the Freedom program had not yet delivered any space hardware. As a result of the cost problem and the perceived lack of maturity, the incoming Clinton administration directed NASA to redesign the Space Station to cut costs. President Clinton formed a Presidential Blue Ribbon Panel led by Charles Vest, the President of the Massachusetts Institute of Technology, to review the result of the NASA redesign efforts and make a recommendation on the Space Station configuration.

The Space Station Redesign Team (SRT). During the winter of 1993, a team of NASA engineers from across the agency was assembled in Crystal City, Virginia, close to NASA Headquarters. The team was charged with identifying Space Station configurations that could be completed at cost estimates that were much lower than the projected \$25 billion of the Freedom program. The team also considered changes in the management structure of the program. In June 1993, the SRT made its final recommendation to the Vest committee. The Vest committee then passed its recommendation on to the President. The SRT had developed technical configurations and cost estimates for three different designs of space stations. Option A was a significant departure from Freedom but used many of the same components. Option B was primarily the Freedom design at that time, and Option C was a dramatic departure using one very large pressurized element, much like the Skylab station. The recommended option was option A, and the Vest committee then referred that option to the President (10).

The Vest committee also took on the task of recommending a new management structure for the Space Station program. The management would now be located at a NASA field center (once again at the JSC); however, it would be in a host relationship where the center did not have ultimate management authority for the program. That authority resided at NASA Headquarters in Washington, D.C. (The program eventually returned to a lead center management structure.) This recommendation forced another difficult transition into the program, but this time at a point when most design activities were complete and many components were being readied for manufacturing. Once again, the Space Station team had to take months to recruit and staff the new office at the JSC. On the industry side, NASA decided that the program must have a single prime contract for the Space Station and that all of the former work package contractors would act as subcontractors to the new prime. Boeing was selected as the prime and began the task of staffing a new Boeing function at the JSC. (During the next few years, Boeing purchased or merged with the other contractors and ultimately formed one contract team.)

In the political world, this marked the absolute nadir of the Space Station program. Although there had been several contentious votes in Congress, the

summer of 1993 was exceptional. An amendment to cancel the Space Station was defeated by a single vote in the House of Representatives. The Senate introduced a similar amendment that was defeated by only a few votes. The Space Station had a new design, new budget, new management structure, and a razor-thin political base.

Russia Joins the Program. In the background of the redesign activities, another NASA team was expanding discussions with the Russians after the breakup of the Soviet Union in 1991. President Bush had initiated an activity between the United States and Russia to fly a U.S. astronaut to the Russian Mir station and to fly a cosmonaut on the Space Shuttle. These tasks were in place and training was begun when the Clinton administration asked NASA to consider engaging the Russians in some way with the Space Station program. A small portion of the NASA redesign team began meeting with Russian engineers to identify components that the Russians could contribute to the new design of the Space Station.

In early 1994, the Space Station International Partners formally invited Russia to join the program. The newly created Russian Space Agency became NASA's interface for all of the activities to be conducted with Russia. As part of the arrangement that now included the Russians in the newly renamed International Space Station (ISS) program, NASA agreed to buy docking mechanisms from Russia and to fly several NASA expeditions on the Mir station.

The technical changes that resulted from adding Russia were dramatic. Several modules were already planned for use in a follow-on Mir 2 program. These modules would form the base of the Russian contribution to the ISS. The Russians would contribute a spacecraft similar to the Mir science modules for the first building block module. This functional cargo block (FGB in Russian) would be the first element launched to the ISS. The next Russian element would be the service module, a duplicate of the Mir core module and the initial living or habitation space for the ISS. The service module allowed the date for permanent crew presence to advance by 18 months compared to the previous plan. Instead of occurring very late in the assembly sequence, the crew would now begin permanent habitation very early in the flow. Russia would also provide logistics support through several flights per year of the Progress spacecraft, and they would provide continuous crew return capability by using Soyuz spacecraft. In addition, Russia planned a number of research laboratories and power augmentation on the Russian segment of the ISS to support these research modules.

Another significant change was that Russia would provide all of the ISS propulsion capability through the service module and the Progress vehicle. This allowed the United States to delete the propulsion capability that was in work. Several years later, the United States would begin development of propulsion capability as a contingency, but those spacecraft were also cancelled.

After the original agreement with Russia was concluded, the United States decided that the critical first element should be a U.S. element. For that reason, the United States contracted through Boeing with Khrunichev, the manufacturer of the FGB. That development proceeded very well and was accomplished on schedule and on budget. Although the United States paid for the module, Russia contributed the Proton launch of the module. The FGB was designed to provide the critical attitude control and propulsion functions for only a few months.

These functions were to be handed off to the Russian service module approximately 5 months later. Due to funding problems, the service module schedule began to slip. To maintain the spacing between the two modules, the United States and Russia also slipped the FGB schedule.

U.S. Development Problems. The early assembly sequence required accomplishing a very complex set of module launches in very close timing. In addition to the FGB and the service module, the U.S. Node 1, the Z1 truss, the first photovoltaic array, and the U.S. laboratory module all had to be completed within just a few months.

Although the Russian schedule was suffering delays, the U.S. development program was also facing problems its own. Both of the two U.S. produced Node structures failed their initial pressure tests. This led to redesign, retesting, and ultimately an agreement with ESA and the Italian Space Agency (ASI) to build two new Node structures. The new nodes would take the place of Node 2 and add an additional Node 3 to the configuration. They would be based on the Multi-Purpose Logistics Modules (MPLM) design and be slightly larger than the initial U.S. Node 1. Node 1 would be modified to increase its strength and would still be the first U.S. produced element in the ISS configuration.

The U.S. design team also needed time to revamp the computer architecture and to finish software development and testing. In addition, several individual systems had problems in final development and testing. While the service module was driving the schedule, the U.S. team had time to fix problems. In addition to problem resolution, a major change was initiated in the U.S. testing program. Several different test scenarios had been proposed for the ISS, but after the 1993 redesign, the NASA team felt that the program did not have schedule allowance for an integrated test of all of the ISS elements together. Obviously, a true integrated test of all of the elements was not going to be possible in any case. Because the service module was driving the schedule, the NASA team could now consider alternative methods of testing and proposed the multielement integrated test (MEIT) scenario. In this plan, the Node 1 and the ground support equipment would be tested together. After the Node was released to launch processing, a Node emulator would then be used to test the Z1 truss, the P6 power module, the laboratory module, the airlock, and the Canadian robotic arm in integrated testing.

MEIT 1 was a great success and significantly increased NASA's confidence that the integrated ISS system would work. NASA and RSA also conducted several tests of all of the computer systems and communication paths in the U.S. lab, the FGB and the service module. All of these tests helped ensure the exceptional performance that was required for the challenging assembly sequence. MEIT 2 was conducted during 2001 to ensure that all of the inner truss segments were fully tested together.

Finally, in late 1998, the ISS team felt that the service module was on a path to completion and that the FGB and Node 1 should be launched. The FGB was launched successfully from the Baikonur Cosmodrome in Kazakhstan in late November 1998. Node 1 was launched on STS-88 in early December just 2 weeks after the FGB (Fig. 3). These early missions were extremely successful, and minor anomalies were easily corrected. Shuttle logistics missions had now been inserted into the sequence before and after the service module. These missions



Figure 3. The FGB mated to Node 1 during STS 88 in December 1998 (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

were eventually needed to conduct repairs and extend the productive life of the FGB because the service module schedule continued to be a problem.

During 1999, the ISS program continued to plan for the launch of the service module. The Russian Proton booster, however, had two failures in 1999, and Russia needed time for failure investigation, recovery, and demonstration of successful Proton flights prior to the SM launch (13). Finally in July of 2000, the service module was launched successfully from the Baikonur Cosmodrome and docked autonomously to the FGB and Node 1 two days later (Fig. 4).

Following the SM launch, the ISS Program completed a rapid succession of flights during the next year. These included three resupply missions, the Z1 truss delivery, the first and second expedition crews, the first U.S. solar array, the U.S. laboratory module—Destiny, the Canadian Space Station Remote Manipulator System (SSRMS)—Canadarm 2, the joint airlock delivery, the first five Progress resupply flights, and the first Soyuz rotation flight (Fig. 5). This flight rate—eight shuttle flights, one Proton flight, two Soyuz flights, and five Progress flights in 14 months is unequalled in human spaceflight experience.

After the SM launch, as NASA now held firm launch schedules, the cost profile of the NASA program started to show troubling indicators. For several years, as the schedule had shifted, NASA had actually underspent the ISS budget plan. Despite earlier cost growth indications (14), the ISS cost profile left the impression that the program could proceed on its cost plan. The new indicators were troublesome and showed that the cost trend in the program was changing. In response to these indicators, NASA took the step of developing a

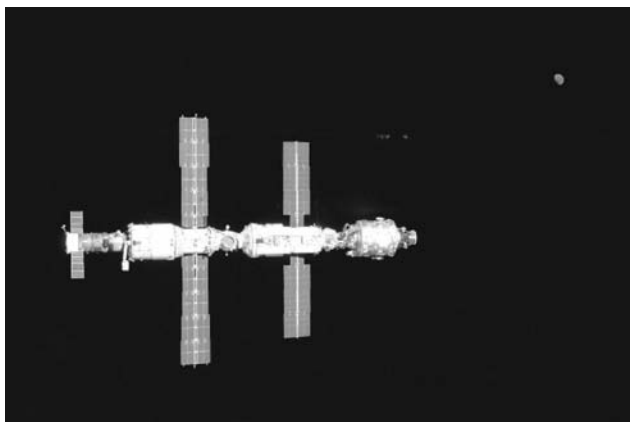


Figure 4. The SM joins the ISS stack in July 2000 (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

new cost estimate for the ISS program. The projected cost growth at the end of the estimating exercise was \$4 billion for the 5-year budget horizon. In response to the budget growth, the incoming Administration terminated the U.S. propulsion module, the U.S. habitation module, and the U.S. crew return vehicle (CRV). In addition, almost \$1 billion was moved from the research budget. The resulting configuration is called U.S. Core Complete and builds through the deployment of Node 2 (Fig. 6). At that point, the International Partner Laboratories can be deployed. The ISS is currently limited to three crew members at a time in this configuration. Throughout 2002, the ISS program has been undergoing intensive review to validate the program baseline and to develop option paths for research-based enhancements over the U.S. Core Complete configuration. This review is expected to continue into 2003.



Figure 5. Expedition 2 crew member, Susan Helms, operates the Canadarm 2 from the Destiny module (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



International Space Station

Elements on orbit (as of July 02)

U.S. core elements to be added

U.S. owned, partner built, to be added

International elements to be added

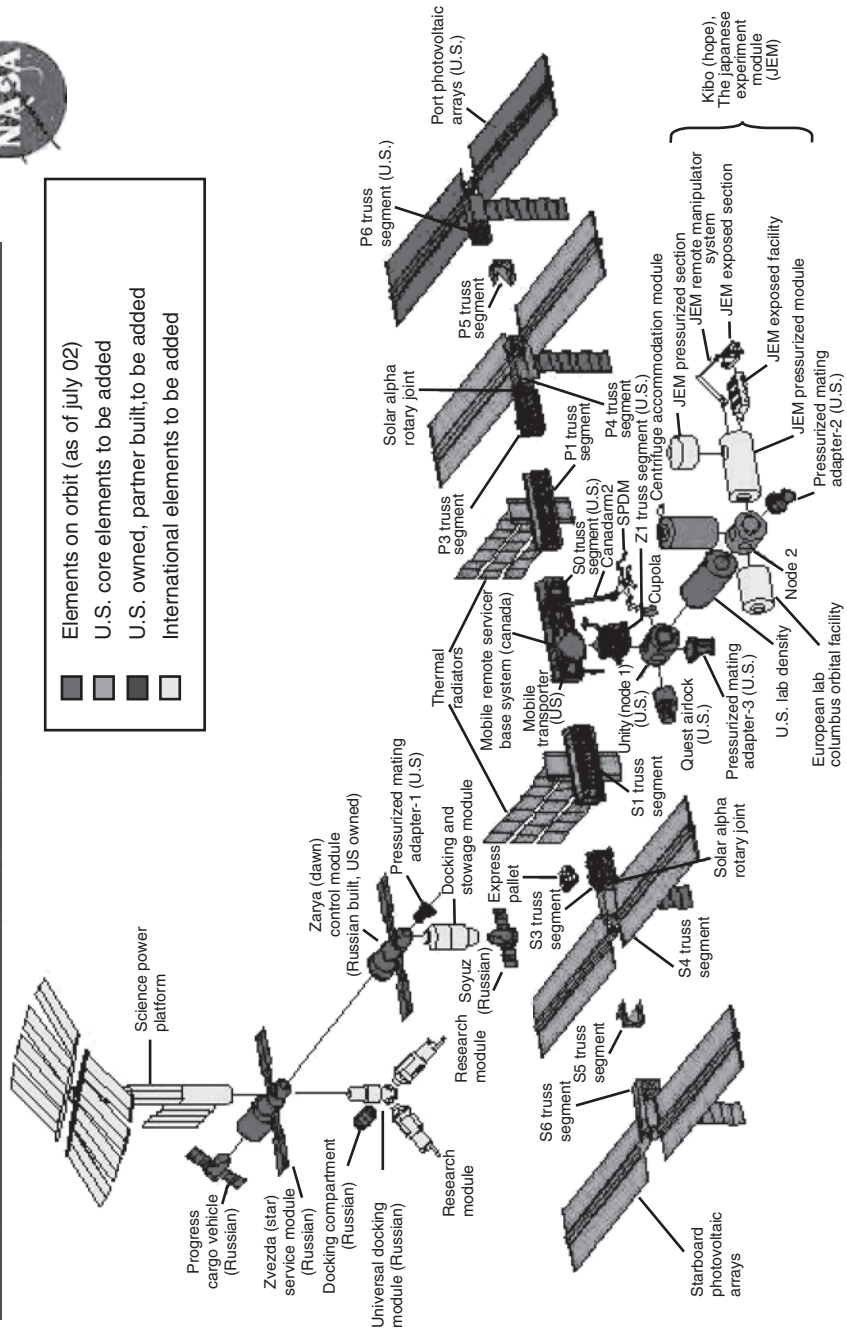


Figure 6. ISS Core Program: President's FY 2003 budget. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The Shuttle–MIR Program

After Russia had been invited to join the new ISS program, NASA decided to expand its participation with Russia by increasing the number of astronaut expeditions on the Mir space station. This plan evolved into a separate program, the Shuttle–Mir Program had its own NASA manager and staff at the JSC. The Shuttle–Mir program also was embraced as part of the overall ISS program and became Phase 1 of the newly formulated ISS program. The principal objective of Phase 1 was learning to work with Russia. The existing International Partners had been a working team since the beginning of the Freedom program. In fact, NASA had a long history of joint projects with ESA, NASDA (the Japanese Space Agency), and CSA. Through the Freedom years, however, NASA and the partners had developed a consistent understanding of engineering standards, practices, and specifications. In contrast, almost no collaboration had taken place with Russia (or the Soviet Union) since the Apollo–Soyuz mission in 1975. Phase 1 proved to be the learning experience that was necessary to bring these two teams together. NASA began with a small team in Moscow to support Norm Thagard's mission on Mir in 1996, but the staffing continued to grow as the involvement of the NASA team in the Mir missions increased. The Shuttle–Mir program demonstrated the ability of the two control teams to work together, the ability of the training system and astronauts to work together, and demonstrated significant technical interfaces through the docking of the Shuttle to the Mir, vehicles of approximately the same mass (15).

NASA's Mir experience also had significant unplanned and traumatic events. During Jerry Linenger's mission, one of the solid fuel oxygen generator (SFOG) canisters caught fire during activation. The fire was successfully extinguished and the crew was safe, but the event provoked both technical and political questions about the remainder of the Phase 1 program (16). Communications between the two teams became very strained during the next several weeks as the Russians worked through the failure analysis of the SFOG canister. In the United States, it prompted additional political oversight, yet NASA chose to continue with the program.

In just a few months, though, the wisdom of continuing was again called into question when the Russians had another dramatic incident on Mir. While the crew was manually docking a Progress resupply ship, the process went awry, and the ship sailed past the docking port striking one of the Mir science modules as it passed the complex. The collision punctured the hull of the Spektr module, and air started to leak from the Mir complex. The crew sealed off the Spektr and fought to regain control of the Mir station. Eventually, the crew used Soyuz thrusters to gain control of Mir and then started the process of charging the batteries and regaining active gyrodyne control of the complex. The U.S. astronaut on board during the collision was Michael Foale. Many of the U.S. teams experiments and many of Foale's personal effects were lost in the Spektr module, and the NASA team had to replan the remainder of the mission around the equipment that Foale could still access.

The Progress collision raised the political oversight of the program to a new high. Congressional hearings, NASA Inspector General studies, and additional external review teams became the norm for the rest of the program. For several

months, NASA provided a daily report of Mir operations to the White House and the Congress. Despite the difficulties, NASA and the external review teams determined that the situation was still safe for flight, and the final two astronaut missions with Dave Wolf and Andy Thomas were completed without incident.

Although the Shuttle–Mir program ended wracked in controversy, it met its principle objectives very well. The events of the program, both good and bad, formed an American/Russian team that learned to work together very well. The other partners had worked together for 10 years during the Freedom program, and Phase 1 of the ISS program brought the Russian team into the partnership. The experience on Mir, however, was based on the US astronauts working essentially as guests on the Russian station. The actual operation of the ISS would require a much bigger step in the overall team and partnership formation.

The ISS Physical Configuration

Overview. The ISS in the current U.S. Core Complete configuration is comprised of pressurized modules, which are contributed by many of the partners, and the external truss elements that distribute resources such as power, cooling, and command and control across the ISS elements (Fig. 7). Each partner is responsible for the design, manufacturing, testing and sustaining engineering of the elements that they provide.

The U.S. Segment. The U.S. segment is comprised of two Nodes, the Destiny laboratory module, the centrifuge accommodation module, the joint airlock, three multi-purpose logistics modules (MPLMs), three pressurized mating adapters (PMA's), a viewing cupola, several pieces of external truss, and the distributed systems of power, thermal control, command and control, and communications. All of the U.S. elements are launched on the Space Shuttle and are therefore designed within the constraints and capabilities of the Shuttle cargo bay.

The Nodes are cylindrical modules approximately 18 ft long and 14 ft in diameter. They have connecting ports at both ends and in four radial locations around the module. Connecting the various modules of the ISS is the Node's principal function; the internal space is used primarily for stowage. Node 1, launched with PMAs 1 and 2 in December 1998, connects the FGB and the Destiny laboratory module (Fig. 8). Nodes 2 and 3 are being manufactured by the European Space Agency (ESA) and ASI as an offset for the Shuttle launch of the ESA Columbus laboratory module. Node 2, scheduled for launch in 2004, will connect to the front of the Destiny module and will provide the connecting point for the European and Japanese segments. In addition, the CAM will mate to the top port of Node 2. The U.S. contribution to Node 3 is currently uncertain due to budget issues. If used, Node 3 will attach to the bottom port of Node 1. PMA 3, launched in October 2000, is currently mounted on the bottom port of Node 1. The locations of PMA 2 and 3 change as a result of assembly operations during the ISS building process.

The Destiny laboratory module was launched in February 2001 and attached to the front port of Node 1 (Fig. 9). Destiny is a cylindrical module 28 ft and 14 ft in diameter that is designed to hold 24 equipment racks of system or research equipment. Ultimately, up to 13 rack locations will be dedicated to

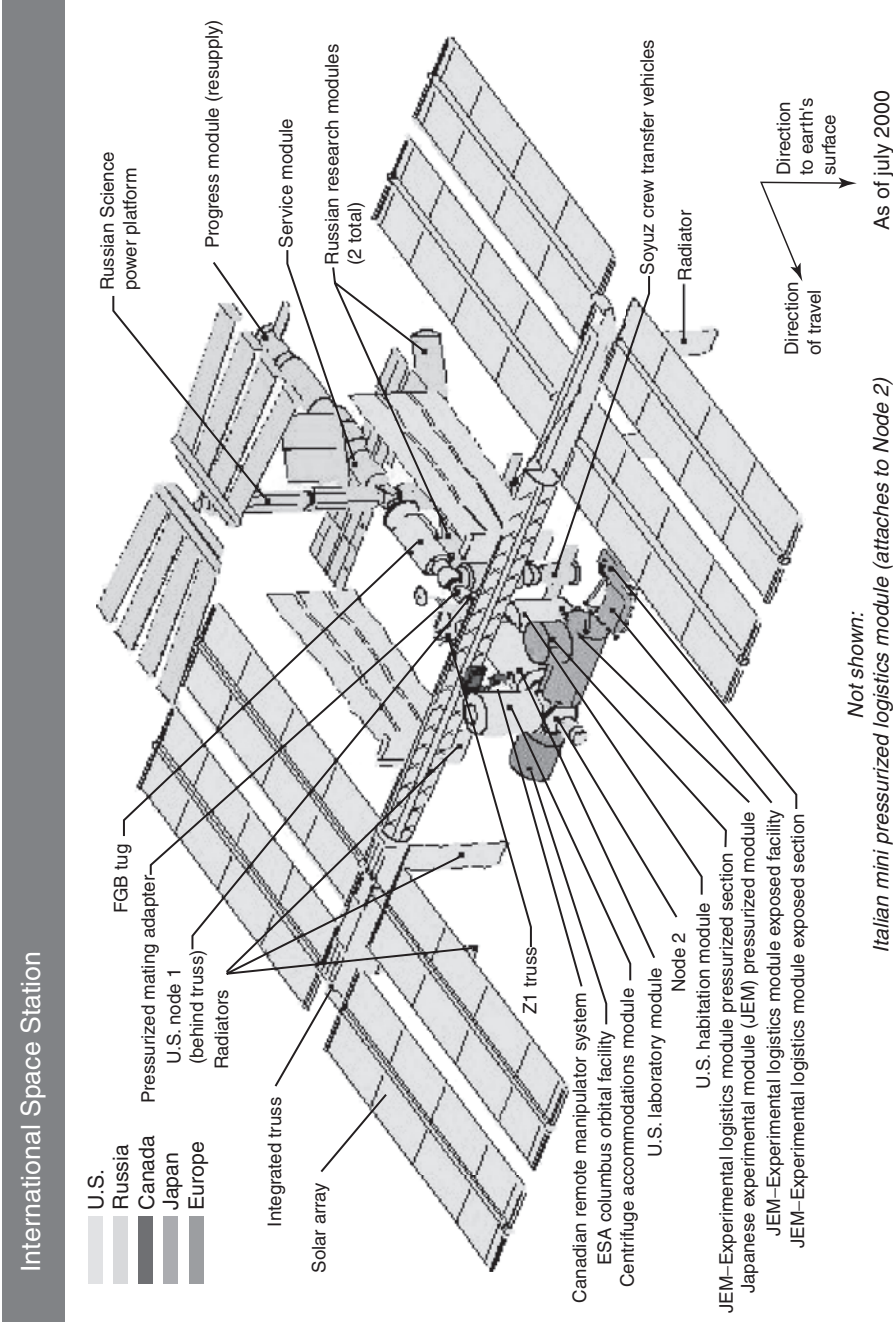


Figure 7. International space station. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

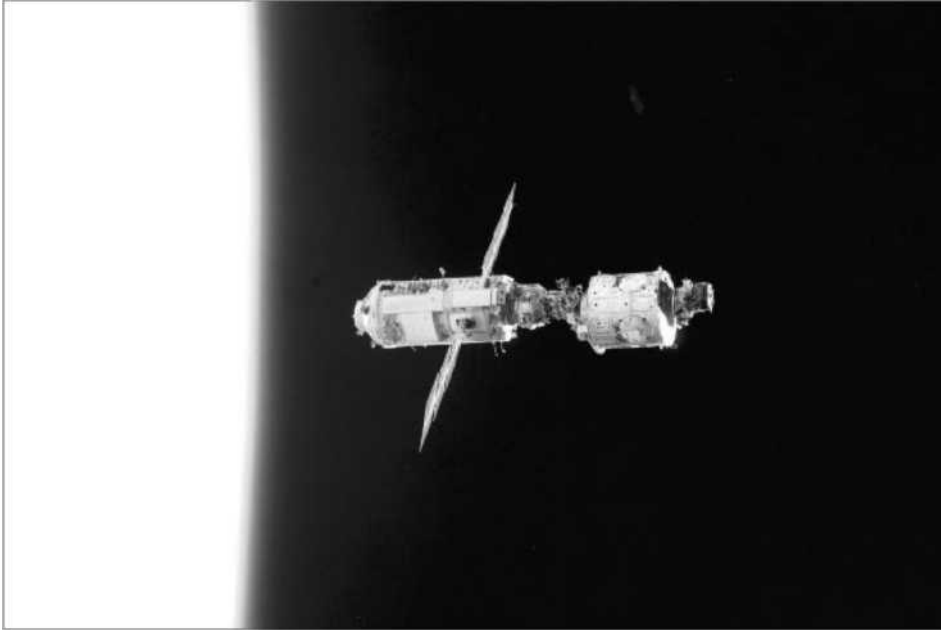


Figure 8. The U.S. Node 1 mated to the FGB with both PMAs 1 and 2 (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

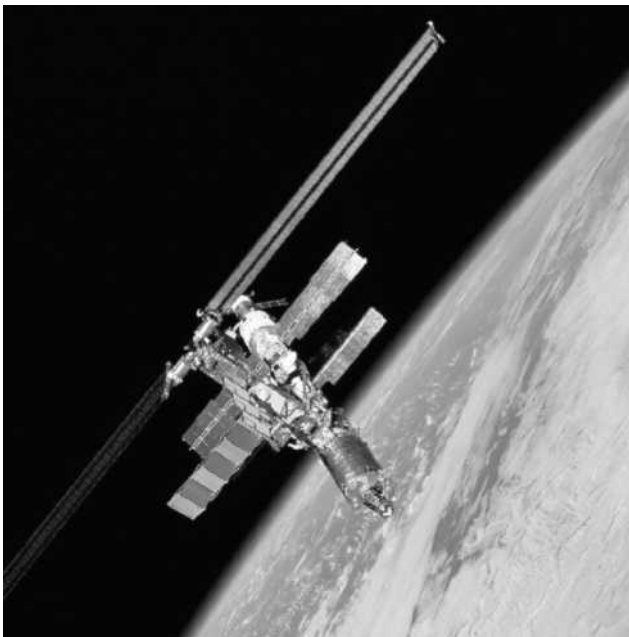


Figure 9. The ISS with the Destiny lab and the first U.S. power module (P6). Note that PMA 1 has moved to the front of the Destiny module (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

research equipment. These racks are launched over time as they are completed. The research racks can also be exchanged over time to allow for new experiment development during the life of the ISS. Destiny also contains the principal U.S. command and control computers and the guidance, navigation, and control (GNC) computers. Another unique feature of Destiny is the presence of an optical quality glass window for use in observation experiments.

ASI built the three MPLMs in exchange for research time and resources on the ISS. All three of the modules—Leonardo, Raffaello, and Donatello—have been delivered to the Kennedy Space Center (KSC). Leonardo and Raffaello have each flown several missions. Donatello will be used in the future. The MPLMs were not designed, however, to stay in orbit. The operational scenario for each flight is to berth the MPLM with the Shuttle or ISS arm to the Node for loading and unloading (Fig. 10). The MPLM is then placed back into the Shuttle cargo bay to return to Earth. The MPLM is configured to carry up to 16 racks worth of equipment and supplies. Due to its connection at the Node, it can use the full 50 in. diameter hatch of the ISS, and this allows the MPLM to transfer the large research racks used on the ISS. The MPLM also allows returning to Earth a large amount of equipment, research samples, and trash.

The joint airlock, Quest, was launched in July 2001 (Fig. 11). It houses the equipment to prepare for space walks. The airlock is compatible with both the U.S. and Russian extra-vehicular activity (EVA) suits. The airlock is divided into an equipment lock and a smaller crew lock to minimize the volume that must be evacuated for each EVA. A special pump captures the ISS air so that not all of the air is vented to the vacuum of space. High-pressure gas (oxygen and nitrogen) to resupply the systems is contained in tanks mounted on the outside of the airlock.

The Z1 truss was launched in October 2000 and mounted to the top or zenith port of Node 1. The truss provides the structural base for the four large control moment gyroscopes (CMGs) and for the communications antennas. The truss also acts as a spacer for the first of the four large solar arrays to be mounted in a temporary location. The first solar power module was launched aboard the



Figure 10. The MPLM in the Shuttle cargo bay approaches the ISS (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



Figure 11. The Quest airlock is mounted to the ISS (NASA photo). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Shuttle in December 2000. It was mounted on top of the Z1 truss to provide power early in the assembly sequence. The power module is comprised of the upper and lower solar array wings, the beta gimbal assemblies, the batteries, a thermal system, and several power conditioning and distribution components. Each solar array is approximately 42 ft wide and 120 ft long. These are the largest solar arrays ever built for spaceflight. Each array, however, folds into a blanket box approximately 15 in. high. The blanket boxes then fold back along the truss for mounting inside the Shuttle cargo bay. The power modules are designed to track the Sun during the ISS orbit by using a beta gimbal assembly about the longitudinal axis and by a large solar alpha rotating joint (SARJ) that rotates around the axis of the primary ISS truss. The SARJ is a very challenging design item and probably the largest rotating joint to fly in space.

The solar arrays provide direct power to the ISS bus system during the sunlit portion of the orbit. They also provide power to the nickel-hydrogen (Ni-H) batteries to maintain their charge. The batteries then provide power during the eclipse portions of the orbit. The system conditions the power and provides for primary and secondary distribution across the truss members and into the modules. The final distribution and control of power is handled within the modules and in many cases at the rack interface itself. The final power at the rack interface is 120 volts direct current (Vdc).

The development of a high-voltage power system for the ISS raised many problems (17). The power standard for previous human spacecraft had been 28 Vdc. All of the components of the ISS power system had development problems due to the large extrapolation from past experience. The ISS itself also presented a large power concern. Due to the large size of the solar arrays, the ISS creates a charged plasma field in the near environment of the ISS. Early on, NASA identified a safety concern that an electric arc could pass from this charged plasma to the truss of the ISS. This would be a potentially lethal situation if an EVA crew member were in the path of that electrical arc! To dissipate the electrical energy, NASA developed and flew a device called a plasma contactor unit (PCU). There

are two PCUs mounted on the ISS which use xenon gas to generate a stream of ions to dissipate the charge.

In due course, all four of the power modules are aligned on the truss axis. The first module, P6 will be relocated from its early position on the Z1 truss to its final outboard position in 2003. Due to the early power configuration, the P6 power module also carries extra thermal radiators. The two extra radiators will also be relocated when P6 is moved to its final location. Together, the power modules will provide about 75 kw of power. Using all of the arrays, this allows an allocation of 45 kw to the research mission. This is a huge increase in power available to research compared to that of NASA's past missions.

In April 2001, the Canadarm 2 was deployed on the ISS by its sibling on the Shuttle, the Canadarm. The Shuttle's robotic arm had been developed by the Canadian Space Agency (CSA) as part of an agreement with NASA in the 1970s. When assessing their potential contributions to the ISS, CSA identified an evolution of the original design, which would allow the ISS crew to perform assembly and maintenance robotically. This arm is unique in that it does not have a fixed shoulder joint. By using symmetrical end effectors at each end, it can reposition itself by "walking" from one grapple fixture to another across the ISS. The Canadarm 2 can also be mounted on its mobile base system (MBS), which rides on the mobile transporter (MT) of the ISS. The MT, launched with the starboard zero (S0) truss segment in April 2002, allows the Canadarm 2 to ride across the front face of the ISS truss. Ultimately, CSA will also provide a highly dexterous robotic end effector for the Canadarm 2 to enhance its maintenance function. The special purpose dexterous manipulator (SPDM) is a two-armed robot that mounts in one of the Canadarm 2 end effectors. From there, it can use its two arms to remove and replace the critical avionics and control boxes that are outside on the ISS truss. Like its predecessor, the Canadarm 2, or Space Station/Remote Manipulator System (SSRMS), has been extremely valuable for assembly and operating tasks. The crew currently controls the SSRMS from stations located in the Destiny lab module.

The next two truss pieces are scheduled to fly in 2002. These two pieces, the port 1 (P1) and starboard 1 (S1) trusses, attach at both ends of the S0 truss. They are essentially mirror image elements whose principal function is to provide the ISS thermal control system. Each of these truss pieces contains three large thermal radiators that extend from the back face of the truss. Like the solar arrays, these radiators rotate to maximize the cooling capability of the ISS. The internal thermal control system uses air and water loops to collect the heat generated inside the modules. This thermal energy is passed to the external loop through heat exchangers. The external thermal loop uses ammonia to transport thermal energy to the large radiators on the P1 and S1 trusses.

The plan for 2003 includes the launch of the P3/4 and S3/4 power modules. These truss pieces connect to P1 and S1 and contain the large SARJ joints. The power module components of these elements are essentially identical to the equipment already in orbit on the P6 truss. Once these elements are in place, the P6 power module will be stowed and relocated at the end of the port side of the truss. The P5 spacer truss is required for mounting P6 with proper clearance on the truss. Once the starboard spacer truss (S5) is complete, the last solar power module, the S6, will be deployed in 2004.

The launch of Node 2 is the key milestone for 2004. Once Node 2 is in place, the European and Japanese labs can be mounted in position. Node 2 is one of the nodes developed by ESA and ASI as compensation for the launch of the ESA laboratory. Its design is slightly larger than the original U.S. Node design. It is designed with interfaces to accommodate the Japanese experiment module (JEM) elements on the port side, the European lab on the starboard side, the centrifuge accommodation module on the top port and the periodic berthing of the MPLMs at the bottom port.

The U.S. segment also contains components of several distributed systems, which provide capability across the ISS. The electrical power and thermal control systems have already been discussed. The other primary systems are the command and data handling (C&DH), telemetry, crew health; life support; and guidance, navigation, and control. The C&DH system consists of the many computer processors across the ISS that control the many functions and distribute data to the crew and the ground team. The computer system is comprised of three top-level command and control computers and many other local computers that controls individual systems areas. The computers are actually multiplexer/demultiplexer (MDM) devices that are outfitted with the primary microprocessors for the ISS. The MDMs are outfitted with a radiation-hardened Intel 386 microprocessor chip. Although several generations old, this chip has the advantage of radiation testing and a large amount of operational experience. The command and control MDMs maintain the overall hierarchy of another 43 MDMs that control other functions. Many of the MDMs are housed within the pressurized volume; however, several are located outside on the truss elements.

The guidance, navigation, and control system is the most distributed system on the ISS. This system, which controls the altitude and orientation of the ISS, must use functions and components across the U.S. and Russian segments to accomplish the mission. The system software is housed in GNC MDMs within the U.S. lab. The software for the GNC takes sensor input from the service module sensors and the Global Positioning System (GPS) to determine the ISS orientation. The MDMs can then command either the US CMGs or the Russian propellant thrusters to adjust the orientation. The Russian thrusters are also used for reboost to maintain altitude.

The life support system is also distributed across the U.S. and Russian segments but not integrated to the level of the GNC. The service module provides oxygen generation and carbon dioxide removal. The Russian system also provides humidity control and converts the condensation to usable water for the oxygen generation system. The Russians also use the solid fuel oxygen generator canisters (SFOG) for oxygen when required. The U.S. system provides additional carbon dioxide removal and also provides oxygen and nitrogen through the gas systems for the airlock. Both the service module and the U.S. lab monitor the air for toxic substances. Air samples are also taken regularly and transported to the ground for analysis.

The principal habitation capability is in the SM: two crew quarters, the galley and wardroom, and the exercise equipment. The exercise equipment consists of a bicycle ergometer and a treadmill. Personal cleansing and waste management are also done through the SM capabilities. One crew sleep station

has been constructed in an empty rack location in the U.S. lab. Ultimately, additional crew stations can be housed in Node 2. The U.S. lab contains the crew health monitoring and checkout equipment. Outfitted with basic diagnostic tools, this facility is essential to monitoring the health of the crew while on the ISS.

The other components of the U.S. segment that had been planned included the cupola, built by ESA, Node 3, a habitation module, regenerative life support equipment, and a crew return vehicle. Due to the budget issues of 2000–2001, these elements are still in question. The ultimate configuration will depend on NASA's success in revamping the ISS program. Low-level critical activities are continuing on the regenerative life support system and Node 3 development.

The Russian Segment. As discussed earlier, the FGB was purchased by the United States and is counted as part of the U.S. segment. Operationally, however, active control and maintenance is done as part of the Russian segment (Fig. 12). The rest of the Russian segment includes the service module, the Pirs docking compartment, that is mounted on the top port of the SM docking compartment and the Soyuz and Progress spacecraft. Future plans call for a docking and stowage module to extend from the bottom of the FGB, a universal docking module to extend from the bottom of the SM, two research modules, and a replacement docking module. A science power platform (SPP) will provide the power required for the Russian segment's research operations. Russia has been considering changes in some of the components that are planned, and some changes should be anticipated. The financial situation in Russia continues to hamper construction of the additional components of their segment.

Except for the SPP, which is planned to be launched on the Space Shuttle, all of the Russian elements will be launched on Russian boosters. The Soyuz crew vehicle and the Progress resupply ship have been used consistently since the launch of the first crew in 2000.



Figure 12. The Russian contribution. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

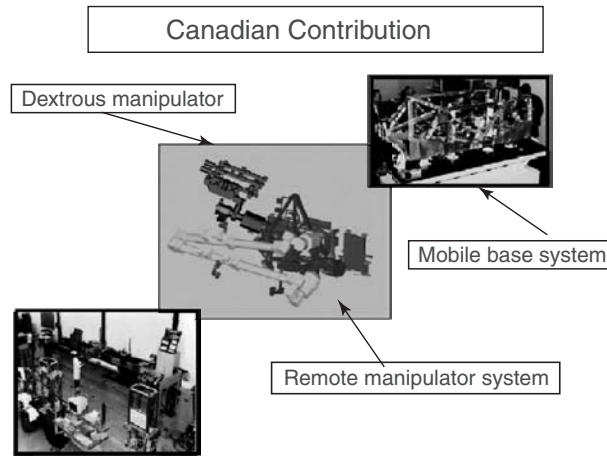


Figure 13. The Canadian contribution. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The Canadian Segment. The Canadian segment has already been described briefly (Fig. 13). The Canadarm 2 was deployed in April 2001. Although it has suffered two failures, it has been extremely useful. In fact, its use has exceeded expectations. The MBS was launched on June 2002. It provides the attachment to the ISS mobile transporter or railway across the front face of the truss. The MBS also acts as the maintenance depot for the Canadarm 2 and is outfitted with additional components. The SPDM is nearing completion and will be available for flight in 2004.

The European Segment. The European segment consists of the Columbus laboratory and the Ariane transfer vehicle (ATV) (Fig. 14). The Columbus is a cylindrical laboratory that will accommodate 10 research racks. It also has external viewing ports that allow viewing or exposure experiments. The Columbus

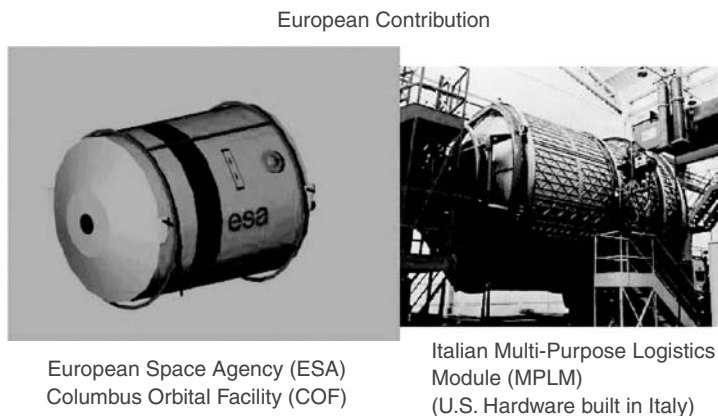


Figure 14. The European contribution. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

is designed to be launched on the Space Shuttle and mated to the starboard side of Node 2. The ATV is designed as a large cargo carrier for the ISS. It will launch on an Ariane booster and dock automatically to the ISS at the rear docking port of the SM. There, it will provide propellant to the SM and will also have its own propellant for ISS use. The ATV will be a significant component of the ISS logistics system and will provide redundancy for the Russian Progress ships. The ATV is expected to be available in 2004.

The Japanese Segment. The Japanese segment is comprised of three elements (Fig. 15). All of these elements are referred to as Kibo, Japanese for “hope.” The three elements are the Japanese experiment module (JEM) laboratory, the JEM exposed facility (EF), and the JEM experiment logistic module (JEM-ELM). The JEM lab is a cylindrical module that can also accommodate 10 research racks. The JEM is unique in that it has a small airlock at the end to allow passing experiments and support equipment from inside the module to the exposed facility outside. The JEM lab module is mounted on the port side of Node 2. The JEM-EF is mounted at the far end of the JEM lab module. It has mounting equipment for 10 external payloads. There is also a remote manipulator system mounted on the back side of the JEM lab for handling the experiments on the JEM-EF. The JEM-ELM is mounted on a special port in the top of the JEM lab module. It will be used for experiment logistics stowage.

Japan is building a transfer vehicle for use with the Japanese launch vehicle, the H2A. The H2A transfer vehicle (HTV) is planned as an additional logistics carrier for the ISS. The HTV is designed to be a dry cargo carrier as opposed to the ATV that carries propellants. The JEM elements are expected to be ready for launch in 2004, but the Japanese schedule is currently under review. The HTV is not expected to be ready until 2006 or later.

Although not part of the Japanese segment, Japan is building the CAM as compensation for the Shuttle launch of the JEM, JEM-EF, and JEM-ELM. The

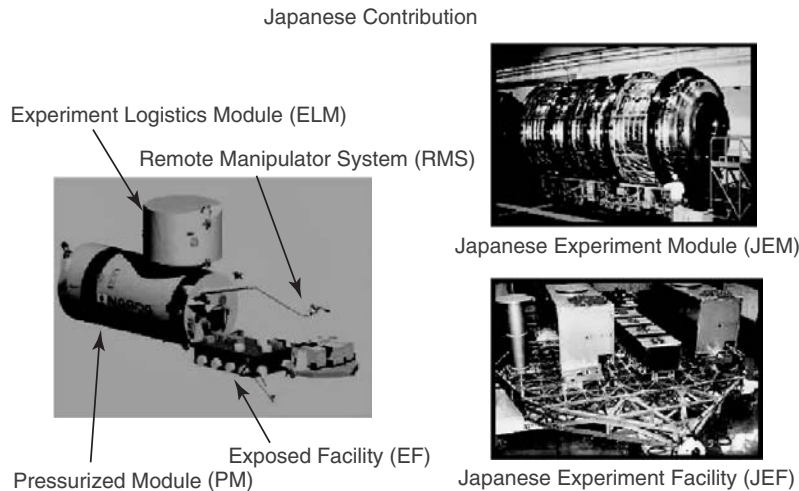


Figure 15. The Japanese contribution. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

CAM is a cylindrical laboratory module designed to house a 2.5-meter diameter centrifuge in its end cone. In addition, the CAM will house a glovebox and several habitat holding racks. The CAM mounts on the top port of Node 2.

The Partner Sharing Formula

Based on the original partnership without Russia, the United States, Canada, Europe, and Japan had agreed on a methodology to share the resources of the Space Station. The sharing agreement is based, in general, on the value of each partner's contribution. This sharing formula is then used to allocate all of the critical ISS resources among the partners. The most important resources include crew time, upmass, and power use. More important than many of these factors, though, is the allocation of opportunities to fly each partners national astronauts. Additionally, each partner would pay a share of a common set of costs for operating the ISS. The common costs would be shared in the same ratio as the resource allocation. The United States, as the major infrastructure provider, would actually have almost half of the Columbus and JEM for its own use.

Once Russia joined the program, several scenarios were discussed to update the previous sharing formula. All of these failed due to the complex nature of assessing the value of the Russian contribution. In the end, NASA reached agreement with the Russian team by agreeing on the value of the resources that passed between the new Russian segment and the rest of the ISS. This is largely driven by the amount of propellant that the ISS uses each year. As a result, Russia had no residual claim on resources in the other segments, and none of the other partners had claims on Russian resources. All partners still have the right to barter among themselves. The major area of concern was still the right to fly national astronauts or cosmonauts. In the new scenario, it was agreed that the initial crew of three would be Russian and U.S. The opportunities would be evenly split, alternating between two Russians and one American and one Russian and two Americans. In this scenario, the other partners did not earn their crew rights until their elements had been deployed and the crew size had been expanded. In the original seven-person scenario, three would be Russian, and the other four would vary according to the sharing formula.

The overall statements of the rights of each of the partners were embodied in the updated Memoranda of Understanding (MOU) with NASA. What was envisioned as a short, simple exercise to update the original documents ended in 1998 with the signature of a new IGA and MOU (18). This simple exercise demanded the energy of a dedicated team from each of the partners in negotiations that lasted nearly 5 years.

Summary

The ISS is well on its way to fulfilling its mission as a world-class research facility. The in-orbit assembly and activation of the ISS has been extraordinary. The assembly plans for 2002, 2003, and 2004 are well understood and have only a small number of options to consider. The vast majority of the U.S. equipment for these upcoming launches is already delivered to the KSC processing facility. The

Canadian SPDM, the ESA Columbus, and the Japanese JEM elements are all nearing completion. As NASA completes its program reform, plans will be developed for the ISS configuration after the partner modules have been launched. The assembly activities during these years will continue to be challenging and will establish new standards for human space exploration.

BIBLIOGRAPHY

1. Messerschmid, E., and R. Bertrand. *Space Stations: Systems and Utilization*. Springer, Berlin and Heidelberg, 1999, pp. 25–30.
2. Messerschmid, E., and R. Bertrand. *Space Stations: Systems and Utilization*. Springer, Berlin and Heidelberg, 1999, p. 240.
3. Messerschmid, E., and R. Bertrand. *Space Stations: Systems and Utilization*. Springer, Berlin and Heidelberg, 1999, p. 243.
4. Messerschmid, E., and R. Bertrand. *Space Stations: Systems and Utilization*. Springer, Berlin and Heidelberg, 1999, pp. 245–247.
5. NASA: Commerce and the International Space Station, KPMG for NASA, Washington, 1999 (can be accessed on the internet at <http://commercial.hq.nasa.gov/rnp/kpmgindex.html>)
6. Mark, H. *The Space Station*. Duke University Press, Durham, NC, 1987, pp. 195–196.
7. Messerschmid, E., and R. Bertrand. *Space Stations: Systems and Utilization*. Springer, Berlin and Heidelberg, 1999, pp. 17–18.
8. Space Station Freedom Media Handbook. NASA, Washington, 1989.
9. Phillips, S. *Summary Report of the NASA Management Study Group*. NASA, Washington, 1986.
10. Vest, C. *Final Report to the President: Advisory Committee on the Redesign of the Space Station*. Washington, 1993.
11. Space Station Operations Task Force Summary Report. NASA, Washington, 1987.
12. Vest, C. *Final Report to the President: Advisory Committee on the Redesign of the Space Station*. Washington, 1993.
13. Russian Rocket Explodes After Liftoff. CNN.com, October 27, 1999.
14. Chabrow, J. *Report of the Cost Assessment and Validation Task Force on the International Space Station*. NASA, Washington, 1998.
15. Stafford, T., and V. Utkin. *Stafford Task Force—Utkin Advisory Expert Council Joint Commission International Space Station Phase 1 Program Joint Final Report*. Washington and Moscow, 1999.
16. Morgan, C. *Shuttle–Mir: The United States and Russia Share History's Highest Stage*. NASA, Houston, 2001, pp. 92–100.
17. Young, T. *Letter report to Wilbur Trafton, Deputy Associate Administrator for Space Flight (Space Station) from the National Research Council's Committee on Space Station*. Washington, 1995.
18. Agreement Among the Government of Canada, Governments of Member States of The European Space Agency, the Government of Japan, the Government of the Russian Federation, and the Government of the United States Concerning Cooperation in the Civil International Space Station. Washington, 1998.

W. MICHAEL HAWES
NASA, Washington
District of Columbia

INTERPLANETARY MEDIUM

In the commonness of life on Earth, most would expect that the composition of space beyond our atmosphere might be planets, comets, asteroids, and an occasional meteor. In reality, the space between the planets and other larger objects of our solar system is richly composed of a variety of complex phenomena, which on Earth drive weather, affect communications, and provide beautiful displays with the aurora.

Long considered an empty void, the vast space between the planets and our Sun is actually filled with a tenuous gas comprised of neutral and ionized particles along with small dust grains. The source of the ionized particles comprising this gas is mostly outflows and outbursts of the Sun. Some of this gas is due to outflow of particles from planets, comets, and asteroids. Finally, some of this gas comes from the infall of particles of gas and dust from the surrounding interstellar space. In this chapter, we will explore the boundaries, composition, sources, and dynamics of the particles filling the interplanetary medium.

The Interplanetary Medium—Inner Boundary

The inner boundary of the interplanetary medium (IPM) is derived from specific models of gases in the outer atmosphere of the Sun called the corona. Because the Sun is an extremely dynamic object in space, the inner boundary of the Sun fluctuates with the modes of solar activities. In a simplistic argument, the boundary between the corona and the IPM can be defined as that point where the solar corona becomes less dense than other constituents of the IPM. This definition becomes too limited, though, when we realize that the interactions of solar plasmas are also governed by local magnetic fields, and hence trapped solar plasma can extend into the IPM significantly beyond the boundaries of the solar corona.

A possible alternative boundary point between the corona and the IPM is the point where the subsonic dynamics of the plasma of the corona transit into the supersonic flow, known as the solar wind. This boundary is best understood by examining the hydrostatic balances of gases that comprise the solar corona. Parker (1) showed in 1959 that the gas comprising the corona must expand due to pressure balances. Because the Sun is in pressure equilibrium, the outward thermal and magnetic pressures balance the gravitational attraction of the mass of the Sun. The solar wind derives from those particles that escape this boundary. The static equilibrium of the solar atmosphere is determined by the balance of gas pressure and solar gravity. Beyond a certain distance from the Sun, gas pressure exceeds gravity, and supersonic outflow ensues—the solar wind. In general, models show that in steady-state conditions, the exterior boundaries of the corona occur at around 1.01 to 10 solar radii depending on the values of the parameters used in solving the equations (2).

The Interplanetary Medium—Outer Boundary: the Heliosphere

The heliosphere is defined as the region that extends from the exterior boundaries of the Sun to the outermost reaches of the influence of the Sun. The

heliosphere is a magnetic bubble formed by the effects of the Sun's magnetic fields as it interacts with local interstellar winds. As the solar wind flows outward, it interacts with the flow of local interstellar wind and with infalling neutral particles and dust grains. Because the solar wind is a supersonic flow, the transition from the heliosphere into the local interstellar medium, it is believed, occurs as a shock called the termination shock. Farther out from the termination shock is the heliopause boundary layer. The termination shock is the backup of the pressure wave that develops from the heliopause boundary itself. It occurs due to the initial "collision" of the plasma that composes the interstellar wind with the magnetic forces due to the Sun. The location of the termination shock and the heliopause varies significantly based on the activity of the Sun. During the solar maximum, the solar wind is weaker so that the external pressure on the heliopause forces the heliosphere to shrink. The most recent, high, solar activity levels give a potential opportunity for the far-flung Voyager I and Voyager II spacecraft to encounter the termination shock not just once but several times. (As of January 1, 2002, the Voyager I spacecraft was approximately 85 AU from the Sun, and the Voyager II spacecraft was approximately 67 AU from the Sun.) The shrinkage and expansion of the heliosphere occur much faster than the outward motion of the spacecraft. If the termination shock is currently shrinking past one or both of the spacecraft, as it expands, years later it will again pass across the spacecraft to be encountered again. There is the possibility of many such encounters that will provide the opportunity to understand the nature of the termination shock and the heliopause in great detail.

Figure 1 provides a graphical model of the magnetic bubble that forms the heliosphere along with the trajectory of the Voyager I and II spacecraft and the Pioneer 10 spacecraft, as well as the newly proposed NASA Interstellar Probe mission. The interstellar wind impacts the bow shock formed at the heliopause. The termination shock is shown within the bow shock. If current models (2) are accurate, the Voyager I and II spacecraft have a good opportunity to encounter the termination shock sometime on or before 2003. It is estimated that the termination shock at that time will occur at approximately 100 AU.

The Interplanetary Medium—Solar Inputs

The primary source of particles in the IPM is from the Sun in the form of the solar wind, coronal mass ejections (CMEs), and solar flares. The general composition of the plasma injected into the IPM is constrained by the composition of the corona. The primary constituents of the solar wind are approximately 95% protons, 4% alpha particles, and 1% minor ions including multiple ionization states of C, O, Si, and Fe. The solar wind also contains electrons in number approximately equal to the ions, and hence the solar wind is considered an electrically neutral plasma. The solar wind contains approximately 1–10 particles per cubic centimeter. The solar wind is a fast stream of particles that leaves the corona at approximately 400 km/s in the ecliptic plane. The velocity of this stream varies significantly and ranges from 300–1000 km/s. At high heliolatitudes (above 45°) during the solar minimum, the solar wind leaves the corona at approximately 800 km/s, again with a very large range of velocities. Figure 2

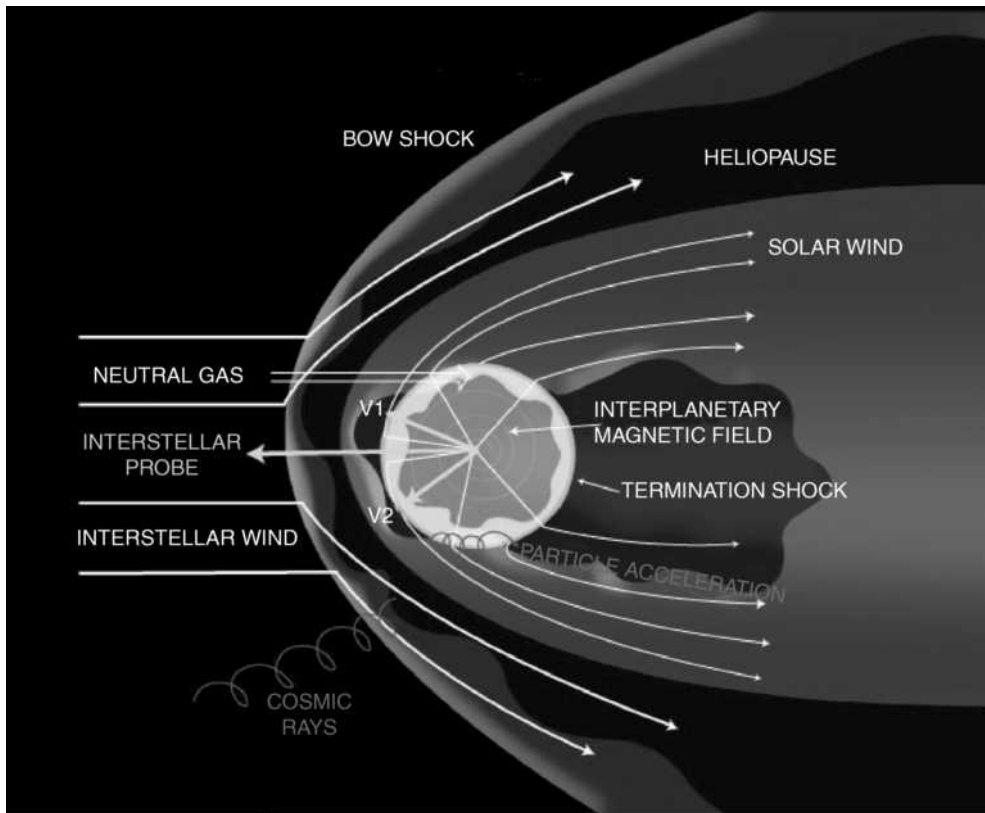


Figure 1. A model of the heliosphere as it interacts with local interstellar winds (adapted from World Wide Web figure courtesy of NASA/JPL). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

shows an image of the Sun (3). Superimposed on this image is a plot of the solar wind speed detected in one polar orbit of the Ulysses spacecraft around the Sun. The figure indicates that within approximately 30° of the ecliptic, the solar wind speed is characterized more by the slower solar wind speeds, although the figure shows that significant variations can occur. Above and below 30° the solar wind speed is characterized more by the faster speeds, again with some variations.

Other solar sources of particles include disturbances in the form of coronal mass ejections (CMEs) and solar flares. CMEs are large-scale bubbles of plasma and embedded magnetic fields that are released from the surface of the Sun. CMEs take hours to develop and are released abruptly. The release of a CME occurs across a large portion of the solar surface and can even affect the entire solar disk. On the other hand, solar flares are smaller scale explosions from the surface of the Sun that take just minutes to form. Solar flares tend to be localized to the area surrounding sunspots. Each of these phenomena transports large amounts of solar plasma into the IPM. The plasma released in a CME or solar flare is more energetic than the steady plasma flow of the solar wind. One last particle type emitted by the Sun is the solar energetic particle (SEP). SEPs are

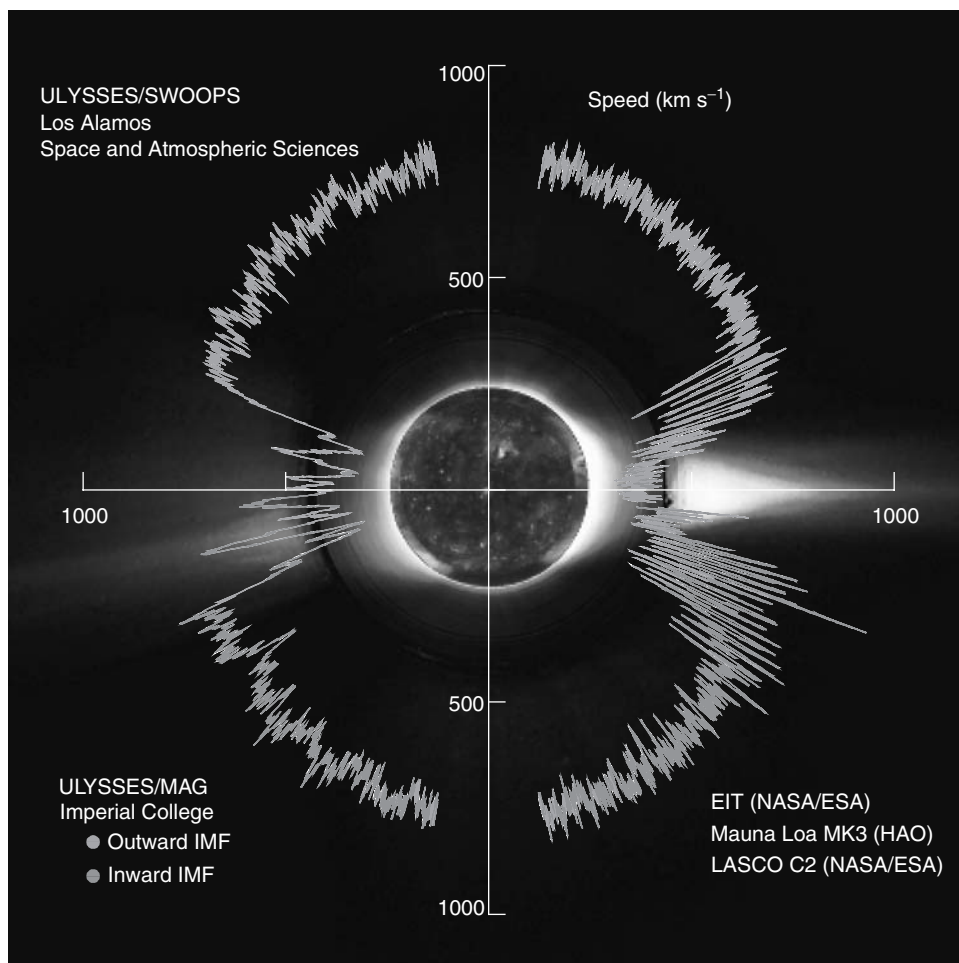


Figure 2. Solar wind speeds superimposed on an image of the Sun [from D.J. McComas et al. *Geophys. Res. Lett.* 25: 1 (1998). Copyright 1998 by the American Geophysical Union]. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

very high-energy ions and electrons that are accelerated by processes within the solar corona, including explosive solar flares. SEP energies are typically between 10 and 100 MeV but can exceed 1 GeV. SEPs provide to solar physicists the opportunity to investigate the composition of the Sun and to understand the accelerative processes that energize the particles.

The Interplanetary Medium—Planetary “Pickup Ions”

Another source of plasma injected into the IPM is due to the interaction of the interplanetary magnetic field with the magnetospheres of magnetized planets. The best examples come from what are called “upstream” particle bursts from

the Jovian magnetosphere (4). Upstream events are characterized by enhancement of ions and electrons, compared to the solar wind. Voyagers I and II, Ulysses, and most recently the Cassini spacecraft have detected short-term enhancements in the overall density of particles as they approached and receded from the Jovian magnetosphere. In general, upstream events seem to occur when the local magnetic field of the IPM is pointed directly toward the planetary magnetosphere. The overall composition of ions during these events is best interpreted as planetary, not solar. The onset of these events is occasionally characterized by the reception of the faster, more energetic ions arriving at the spacecraft before the slower ions. In general, though, the detection of particles during an event occurs at the same time, indicating that the spacecraft is passing through a “flux tube” of magnetically contained plasma that connects directly to the planetary magnetosphere. It is assumed that the magnetic fields associated with these flux tubes have become directly connected to the magnetic fields of the IPM and hence allow for transport of particles away from the planetary magnetosphere.

Figure 3 shows an example of an upstream event that occurred as the Cassini spacecraft was leaving the Jovian magnetosphere on day 37 of 2001. The figure is derived from data taken from the Cassini MIMI LEMMS (Low-Energy Magnetospheric Measurements System) detector and is plotted in distance from Jupiter in Jovian radii. At a distance of approximately 526.7 R_J (post-Jovian

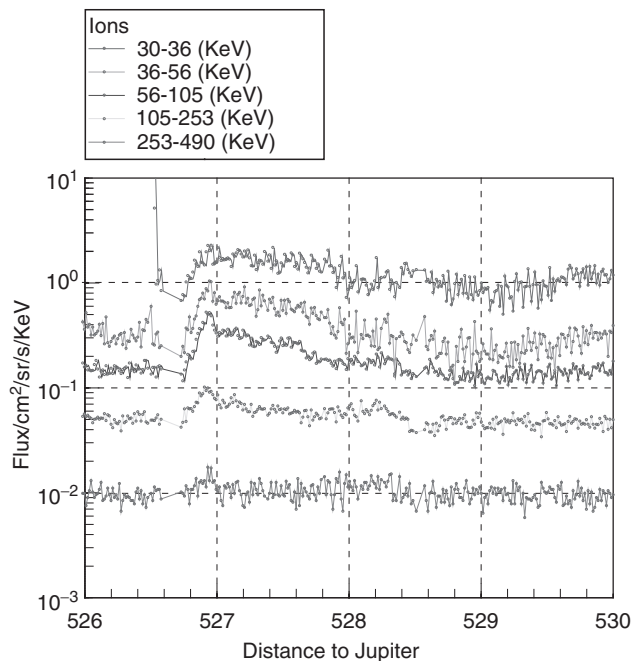


Figure 3. An upstream ion event on the outbound pass of the Cassini spacecraft at a distance of approximately 527 R_J (courtesy NASA/APL, Max Planck Institute, and Fundamental Technologies, LLC). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

encounter—day 37 of 2001), an enhancement of the ion particle flux was noted. In this particular enhancement, the beginning of the event occurred in all of the lower ion energy channels at the same time. This is consistent with an explanation that the spacecraft flew through a magnetic flux tube that is connected to the Jovian magnetosphere on one end. The event lasted for approximately 1 day, and there was notable decay of ion flux in all but the lowest energy channel.

The Interplanetary Medium—Interplanetary Magnetic Field

The other major component of the IPM is the interplanetary magnetic field (IMF) that fills the entire region of the heliosphere. The IMF is mostly due to the transport of solar magnetic structures into the IPM. As the outward moving plasma of the solar wind goes from subsonic in the corona to supersonic just outside the corona, the magnetic field becomes locked within the solar wind plasma. The solar wind then carries the coronal magnetic fields into the IPM. Parker (1) first developed a model of the way the coronal magnetic field is carried into the IPM. The overall model takes into account the shape of the magnetic field lines as the solar wind velocity and the rotation of the Sun determine them. If one assumes a purely radial solar wind and then considers this solar wind in the reference frame of the rotating Sun, then the solar wind has the following components:

$$U_r = u_{\text{sw}}; \quad U_\phi = -\Omega_\odot r \sin \theta; \quad U_\theta = 0, \quad (1)$$

where u_{sw} is the radial solar wind speed, Ω_\odot is the solar angular velocity, r is the distance from the sun, and θ is the solar latitude. This solution assumes that the solar rotational rate on the surface of the Sun is constant instead of the observed latitudinal differential rotational rate. The solar magnetic field lines then follow velocity streamlines. The path is defined by the following equation:

$$\frac{1}{r} \frac{dr}{d\phi} = \frac{U_r}{U_\phi} = \frac{-u_{\text{sw}}}{\Omega_\odot r \sin \theta}. \quad (2)$$

As discussed in the first section, as the solar wind escapes from the solar surface, the solar wind speed becomes a constant at a critical distance from the solar surface. Using this fact, we can integrate equation 4 and derive the following solution:

$$\Delta R = r - R_\odot = \frac{-u_{\text{sw}}}{\Omega_\odot \sin \theta} (\phi - \phi_\odot), \quad (3)$$

where ϕ_\odot is the initial longitude where the solar wind developed on the surface of the Sun. This equation describes an Archimedean spiral where the magnetic field follows a spiraling path as it moves away from the Sun. For low latitudes (slow moving solar wind), the value of ΔR is approximately 6 AU — 1 AU farther than the orbit of Jupiter. For higher latitudes (fast moving solar wind), the value of ΔR is approximately 12–2 AU farther than the orbit of Saturn.

The Interplanetary Medium—Corotating Interaction Regions and Shocks

Based on the combined understanding of the latitudinal solar wind dependency coupled with the observed latitudinal differential rotational rate of the surface of the Sun, the overall structure of the IMF becomes complex. In specific regions, on the boundaries between fast moving solar wind and slow moving solar wind, interplanetary spacecraft have observed magnetically complex structures known as corotating interaction regions (CIRs). CIRs most generally occur when the fast moving solar wind overtakes the slower moving solar wind. CIRs are bounded by two shocks at the edges, called forward and reverse shocks. These shocks are characterized by changes in the magnetic field intensity, the particle density, and the overall magnetic pressure. The shocks associated with CIRs provide an explanation for energetic streams of ions propagating from interactive regions (5).

The effects are best understood by studying how shocks accelerate particles. Magnetized shocks in general are quite efficient in accelerating charged particles and producing an overall increase in the number of energetic ions and electrons in the local plasma environment. Acceleration of charged particles can occur through one of two processes: shock drift acceleration and diffusive shock acceleration. Shock drift acceleration occurs when a charged particle encounters a magnetized structure so that particle acceleration occurs due to induced electric fields parallel to the surface of the shock. The general understanding of shock drift acceleration comes from examining the motion of a charged particle as it encounters a shock. Because the magnetic field strength is larger inside the shock than outside, the gyroradius of a charged particle that impacts a shock decreases. In general, the motion of a particle includes a component of velocity parallel to the shock boundary. In the reference frame of the shock, the particle experiences an electric field parallel to the shock boundary that can either accelerate or decelerate the particle. Acceleration occurs while the particle is outside the shock, and deceleration occurs while the particle is within the shock. Because the gyroradius of the particle is larger outside the shock than inside the shock, the particle spends more time accelerating than decelerating. The factor that determines whether the particle is transmitted through the shock or reflected is based on conservation of the first adiabatic invariant. The condition for reflection is best seen from the following equation that describes the magnetic moment in terms of the pitch angle:

$$\frac{p_1^2 \sin^2 \alpha_1}{B_1} = \frac{p_2^2 \sin^2 \alpha_2}{B_2}, \quad (4)$$

where subscript 1 refers to the conditions outside the shock, subscript 2 refers to the conditions inside the shock, p is the momentum of the particle, and α is the pitch angle of the particle. Under the condition

$$\sin^2 \alpha_1 > \frac{B_1}{B_2}, \quad (5)$$

conservation of the first adiabatic invariant requires that $\sin \alpha_2 > 1$, which is not possible. This condition indicates that a particle is reflected from the shock boundary. If the condition is not true, then, the particle is transmitted through the shock. In either case, though, the motion of the particle includes acceleration while the particle is within the shock. In general, shock drift acceleration energizes a particle by a factor no larger than 10. However, repeated encounters with a particular shock and/or encounters with multiple shocks can allow the energy of a particle to increase dramatically. Another mitigating factor is the relationship between the direction of the shock normal and the direction of the magnetic field. If these are parallel, then shock drift acceleration is not effective in accelerating particles, but if these are perpendicular, shock drift acceleration is very effective in increasing a particle's energy.

Diffusive shock acceleration occurs when a particle encounters a shock that is approaching. By examining the motion of the particle in the rest frame of the shock, it is seen that the energy of the particle is increased. Denote the velocity of the particle before and after the collision with the shock as v_1 and v_2 , respectively. The velocity of the particle in the rest frame of the shock is indicated by primes:

$$v'_1 = v_1 - u_{\text{shock}}, \quad v'_2 = v_2 - u_{\text{shock}}. \quad (6)$$

In the rest frame, the velocity of the particle is simply reflected:

$$v'_1 = -v'_2. \quad (7)$$

The change in energy is then given by combining Eqs. 6 and 7:

$$E_2 - E_1 = \frac{1}{2} m (v_2^2 - v_1^2) = 2m (u_{\text{shock}}^2 - v_1 u_{\text{shock}}). \quad (8)$$

Depending on the value of $v_1 u_{\text{shock}}$ the particle either gains or loses energy. If $v_1 u_{\text{shock}} < 0$, then the particle gains energy; if $v_1 u_{\text{shock}} > u_{\text{shock}}^2$, then the particle loses energy.

When a particle encounters the shock multiple times (due to reflection from other magnetic anomalies outside of the shock), then, the particle can experience repeated acceleration. If a distribution of particles encounters a shock, individual collisions with the shock will be stochastic so that some of the particles are accelerated and some are decelerated. This process has a tendency to spread the particle velocity distribution function to include more slower and faster moving particles.

The Interplanetary Medium—Interstellar Sources of Particles: Pickup Ions

One final source of particles that comprise the constituents of the IPM is the infall of neutral and charged particles from the local interstellar medium. Because the heliopause is a shock boundary, it is difficult for charged particles to penetrate into the IPM. The main component of interstellar particles within the

IPM is neutral atoms. Because these particles are neutral, magnetic fields cannot deflect the motion of these particles. As these particles fall farther into the IPM toward the Sun, solar radiation ionizes them. As the particles become ionized, they become bound to the IPM's magnetic field lines, and their overall motion becomes trapped within the outflowing solar wind. Blum and Fahr (6) originally proposed the concept of these particles, but they were not discovered until the Ulysses spacecraft entered into the quiet regions of the high latitude solar wind. The SWICS instrument (Solar Wind Ion Composition Spectrometer) of the Ulysses spacecraft provides the capability of capturing *in situ* distribution functions. A key discovery occurred when the SWICS instrument found an anomalous component of the solar wind distribution function. Figure 4 shows the distribution function from the SWICS instrument (7). The phase space density is shown as the dotted curve that has a maximum peak velocity of the solar wind velocity. Instead of the expected phase space density, the SWICS instrument showed that the local IPM was filled by a broad distribution of particles that form a slowing declining plateau in velocity space. This plateau, it is understood, indicates the existence of ionized interstellar hydrogen. The falloff of particles whose velocities are greater than $2 v_{sw}$ is an expected result due to the filling-in nature of the pickup ions in velocity space. In the solar wind rest frame, the pickup ions are seen as an isotropic spherical shell whose radius is $1 v_{sw}$. When transformed into the rest frame of the detecting spacecraft, this filled-in shell appears as a distribution of particles whose velocities are between 0 and $2 v_{sw}$. Most interstellar neutrals are easily ionized by the solar radiative source. These particles penetrate to within 6 AU of the Sun. Because helium is more difficult to ionize than all other interstellar neutrals, helium penetrates more closely to the Sun at a cutoff distance of approximately 4.82 AU.

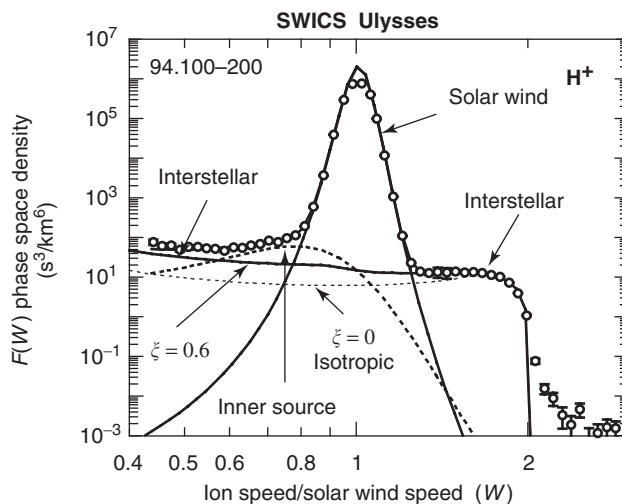


Figure 4. SWICS distribution function showing the 800-m/s solar wind but also indicating an unexpected plateau of particles indicating proton pickup ions [from G. Gloeckler and J. Geiss. *Interstellar and inner source pickup ions observed with SWICS on Ulysses. Space Sci. Rev.* 86: 127 (1998), Fig. 2. With kind permission of Kluwer Academic Publishers].

Conclusion

Existing NASA satellite programs continue to provide a significant amount of data regarding the nature of the constituents of the interplanetary medium and transport processes. Proposed programs such as the Interstellar Probe (trajectory shown in Fig. 1) represent significant opportunities to clarify further the details of the overall structure of the heliosphere and provide *in situ* measurements of the plasma constituents within the interplanetary medium. To continue developing our understanding of the interplanetary medium and the overall heliosphere, we must maintain a continued presence in space.

BIBLIOGRAPHY

1. Parker, E.N. Extension of the solar corona into interplanetary space. *J. Geophys. Res.* 64: 1675 (1959); Bame, S.J., B.E. Goldstein, J.T. Gosling, J.W. Harvey, D.J. McComas, M. Neugebauer, and J.L. Phillips. Ulysses observations of a recurrent high-speed stream and the heliomagnetic streamer belt. *Geophys. Res. Lett.* 20: 2323 (1993).
2. Cravens, T.E. *Physics of Solar System Plasmas*. Cambridge University Press, Cambridge, 1997.
3. McComas, D.J., S.J. Bame, B.L. Barraclough, W.C. Feldman, H.O. Funsten, J.T. Gosling, P. Riley, R. Skoug, A. Balogh, R. Forsyth, B.E. Goldstein, and M. Neugebauer. Ulysses' return to the slow solar wind. *Geophys. Res. Lett.* 25 (1): 1 (1998).
4. Krimigis, S.M. Observations of energetic ions and electrons at interplanetary shocks and upstream of planetary bow shocks by the Voyager spacecraft. *Proc. Int. Symp. Collisionless Shocks*, Hungary, edited by K. Szego, 1987, p. 3.
5. Sanderson, T.R., R.G. Marsden, K.-P. Wenzel, A. Balogh, R.J. Forsyth, and B.E. Goldstein. Ulysses high-latitude observations of ions accelerated by co-rotating interaction regions. *Geophys. Res. Lett.* 21: 1113 (1994).
6. Blum, P.W., and H.J. Fahr. Interaction between interstellar hydrogen and the solar wind. *Astron. Astrophys.* 4: 280 (1970).
7. Gloeckler, G., J. Geiss, E.C. Roelof, L.A. Fisk, F.M. Ipavich, K.W. Ogilvie, L.J. Lanzerotti, R. von Stieger, and B. Wilken. Acceleration of interstellar pickup ions in the disturbed solar wind observed on Ulysses. *J. Geophys. Res.* 99: 17,637 (1994).

JERRY W. MANWEILER
Fundamental Technologies LLC,
Lawrence, Kansas

J

JUPITER

Jupiter is a beautiful and inspiring planet and has returned major scientific dividends on every investment in exploring it. This largest planet in the solar system has more mass than all of the other planets combined, raging storms that last for decades and longer, 16 moons, the largest bigger than Mercury, and a magnetosphere so vast it would appear as big as the full Moon in the sky—Saturn routinely passes through Jupiter’s magnetotail. There are several reasons why Jupiter gets the lion’s share of attention in planetary exploration. First, it is an extremely common planet type. For example, it is likely that many solar systems have a dominant gas giant planet whose orbital radius is in the neighborhood of the “snow line,” the approximately 5 AU distance from a protosun where the temperature in the protoplanetary nebula drops to the point that ice becomes stable. Its position makes Jupiter the protector of the inner solar system. Because Jupiter contains most of the solar system’s planetary mass, accurate observations of its composition, in particular, the hydrogen-to-helium ratio and the precise abundances of elements heavier than helium, are crucial to the success of solar-system formation models and serve as a springboard for cosmochemistry. Jupiter is the case study for all gas giants, including Saturn, Uranus, Neptune, and the nearly 100 extrasolar gas giants discovered to date. Second, the meteorology that shapes the planet’s appearance is more predictable than Earth’s weather. Studying the differences and similarities has led to a better understanding of atmospheric dynamics and chemistry in general. Third, Jupiter’s magnetosphere is second only to the Sun’s; it has an active aurora and strong interactions with its satellites and the solar wind. We can expect that someday a high-order spherical harmonic mapping of the magnetosphere will be used to probe the planet’s interior structure. Finally, because Jupiter is the closest gas giant to Earth, it is the easiest to observe, and because every probe headed for the outer solar system uses Jupiter for a gravity assist, it has had the

largest number of close encounters with spacecraft, such that every investigation of the Jovian system is leveraged by its many predecessors (1).

Atmosphere

Jupiter is instantly recognizable by its bands of clouds and its large, ruddy eye, the Great Red Spot. In 1979, the twin Voyager spacecraft returned the first crystal-clear images of the planet, revealing huge stable ovals standing alongside filamentary swirls of white, red, and brown that are as striking to the first-time viewer as any masterpiece of artwork. Jupiter has six times as many jet streams as Earth and storm systems that last a lifetime or longer. In all respects it is a giant planet; it has 71% of the solar system's total planetary mass and more than 120 times the surface area of Earth.

Studies of the dynamics of Jupiter's atmosphere show that it has just as much in common with Earth's oceans as it does with Earth's atmosphere. Dynamically, Earth's oceans are giant in size compared to the size of ocean eddies—an ocean basin spans hundreds of eddies—just as Jupiter's clouds play host to hundreds of storms. In contrast, Earth's atmosphere is a crowded place and has room for only about a half-dozen large eddies at once. Earth's atmosphere receives more energy per area than any other planetary atmosphere in the solar system, even more than Venus, which is closer to the Sun but more reflective; so perhaps it is no coincidence that our home planet turns out to have the most unpredictable weather in the solar system. But, Earth also has the weakest winds of all of the planets, and so we must guard against making overly simplistic statements about the role of sunlight. As a case in point, the jet streams on Jupiter are more than twice as fast as those on Earth, even though the sunlight received there is 25 times weaker. And, the winds on the other gas giants are even faster than on Jupiter, even though the Sun is just a bright star in the sky as seen from Saturn, Uranus, and Neptune.

The lack of connection between sunlight and wind speed is a major result of the space-craft era, and to it we can add the following two lessons from Earth's air-sea system when considering Jupiter's winds. First, the currents in Earth's oceans are mostly caused by wind action—they would run down if the wind stopped blowing. Likewise, Jupiter may have extremely deep, even powerful circulations that are spun up from the alternating jets overhead, rather than the other way around, a tail-wags-the-dog viewpoint that, nevertheless, comes naturally to oceanographers. Second, the process of creating a wind-driven ocean current does not involve a one-dimensional push of the water by the wind, like a child pushing a toy car. Rather, it is a three-dimensional Coriolis reaction to tilts of the constant-pressure surfaces in the top layers of the ocean (caused by Ekman pumping, a boundary-layer effect). The main lesson is that it is not hard to generate strong, alternating jet streams on a rapidly rotating planet, so long as it is not hard to cause and maintain tilts of the constant-pressure surfaces. Considering that planetary rotation is a vast reservoir of momentum and sunlight is not required to keep the planets rotating, perhaps we should not be surprised after all to see that sunlight is not directly involved in determining wind speed. Instead, we should focus on which processes, including moist convection, internal

heat, and sunlight, act to warp constant-pressure surfaces in an atmosphere and let the Coriolis effect take care of the rest.

Observations. Jupiter appears to the naked eye as a large, bright star; an ordinary pair of binoculars reveals its famous banded clouds, its Great Red Spot, and its four largest moons, called the Galilean satellites, which Galileo discovered 4 centuries ago using one of the first telescopes. The dark bands on Jupiter are traditionally called belts (dark like a belt worn around a waist), and the light bands are called zones. For a quarter-century, high-resolution images from the spacecrafts Voyager, Galileo, Cassini, and the Hubble Space Telescope, have revealed fine details in Jupiter's clouds (Fig. 1) that allow determining the winds reliably and repeatedly. Jupiter is particularly photogenic; in contrast, cloud tracking is not reliable for deducing wind speeds on Earth, Uranus, or Neptune, because their clouds are more ephemeral than on Jupiter; Saturn is better, except that its cloud features are muted by an overlying haze. The peaks of Jupiter's jets are centered precisely on the abrupt color boundaries between the belts and zones, and the speed and location of the jet streams have shown virtually no change during the spacecraft era, even though the color contrast of the jets often changes. The exception that proves the rule is the strong eastward jet at 23°N , which has shown a 20% variation in speed during the spacecraft era; most of the other jets have shown no variation at all.

Jupiter's surface clouds top out at a pressure level of about 0.7 bar (1 bar = 10^5 Pa, approximately sea-level pressure on Earth); the tops in the zones

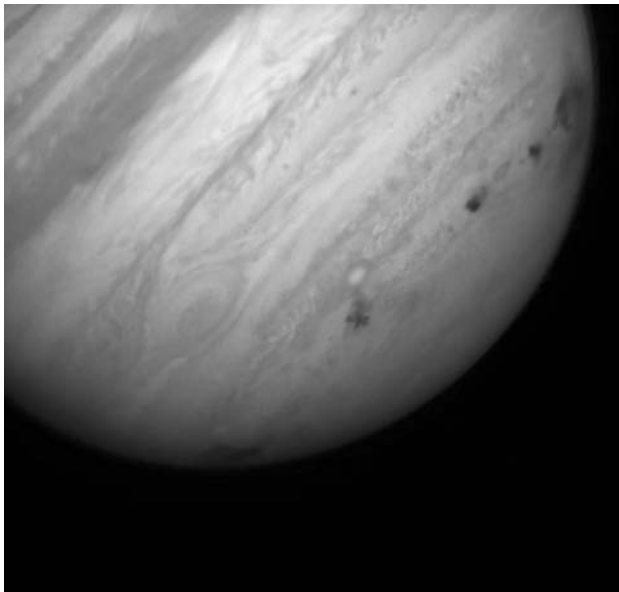


Figure 1. Hubble Space Telescope (HST) image of Jupiter bearing the scars of the Comet Shoemaker–Levy 9 impacts. From left to right, the impact sites are the E/F complex on the edge of the planet, the star-shaped H site, the sites for N, Q1, Q2, and R, and the D/G complex on the far right limb. Also visible is the Great Red Spot and several smaller storms (Hubble Space Telescope/NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

are somewhat higher than in the belts. The clouds are predominantly white ammonia ice that is colored by unknown chromophores, probably sulfur or phosphorus compounds in trace amounts; the chemical makeup of the chromophores has been a long-standing open question that will probably require *in situ* measurements to answer. Occasionally, a little red spot appears in the Northern Hemisphere that has spectral properties identical to those of the Southern Hemisphere's Great Red Spot and at about the same latitude reflected across the equator. Why this particular latitude range, 19–23°, consistently harbors the reddest pigment found in the atmosphere is an intriguing question. The Galileo Orbiter has provided evidence that not all of the surface clouds are ammonia. Specifically, the occasional, sudden appearance of a bright-white cloud seen in the filamentary-cloud areas with cyclonic shear ("cyclonic" meaning in the same direction as the planet's rotation, implying low pressure), is the top of a giant cumulus *water* cloud that has punched up through the ammonia cloud from the 5–6 bar region below. These water clouds stand about three times taller than cumulus towers on Earth and would be breathtaking to view from the side; the largest ones erupt on the northwest side of the Great Red Spot and cover an area the size of Alaska before they are sheared apart. Such energetic moist convection may hold the key to the transfer of energy from Jupiter's interior to its atmosphere.

Jupiter has hundreds of stable oval-shaped storms (Fig. 2) that populate its anticyclonic (high-pressure) shear zones. They tend to get more numerous and smaller as latitude increases toward the poles. The Cassini flyby data show that the alternating jets extend poleward all the way to at least $\pm 80^\circ$, even though the belt-zone banding gives way to a leopard-skin appearance poleward of 60° . Ultraviolet-filter images that are sensitive to aerosols in Jupiter's stratosphere

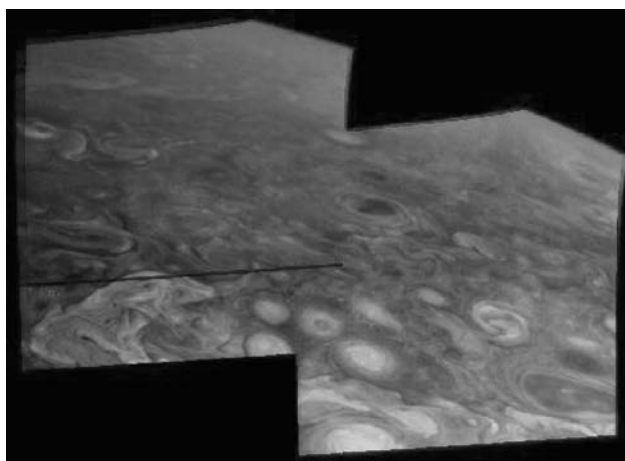


Figure 2. Galileo spacecraft image of Jupiter's swirling storms and stratospheric haze in the vicinity of latitude 50°N . This is a false-color image in which red indicates high-altitude features and blue indicates low-altitude features. North is up and the line of sight looking toward the limb emphasizes the high-altitude haze (Galileo Project/NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

show occasional large vortices and wavy boundaries in Jupiter's stratosphere that are unrelated to the eddies in the troposphere below. Eddies in Jupiter's atmosphere last much longer than their cousins on Earth, but they do have a life cycle, albeit with a life span closer to three score and ten than the couple of weeks for which major storms remain coherent on Earth. For example, at 33°S, the three White Ovals BC, DE, and FA that emerged from a wavy disturbance in 1939 (which had six pinched areas labeled A through F) existed virtually unchanged for 60 years, until they merged together sequentially in the late 1990s to form a single White Oval, called BA. Like Old Faithful, Jupiter's Great Red Spot (GRS) seems to have been a part of the planet's landscape forever—Robert Hooke reported sighting a giant spot on Jupiter more than 3 centuries ago. We can see firsthand how this giant storm persists in the face of dissipation by watching movies of Jupiter's cloud-top dynamics. They show that the GRS ingests a steady diet of small anticyclones that are born in the cyclonic, filamentary region to its northwest and then travel at latitude 19°S all the way around the planet until they are swept into the GRS from its eastern side. This means that the GRS should not be treated as a local storm but as the most obvious component of a global system in dynamic equilibrium, on which the Sun never sets. It would also be a mistake to treat the GRS as eternal; it has been shrinking in longitudinal extent during the past century. The trend has been precisely measured during the last quarter-century and raises the possibility that the GRS may lose its stable elliptical shape and cease to exist at some point, perhaps within the reader's lifetime.

Above the Clouds. Planetary scientists are always hunting for ways to infer vertical information from horizontal (map-view) information that is gathered remotely. Given a rapidly rotating planet like Jupiter, one such important technique allows for the determination of wind shear with respect to height above the cloud tops, where, by definition, there are no visible tracers of the motion. The thermal-wind equation makes this possible; it is a three-dimensional relationship that holds when the Coriolis acceleration strongly couples the atmospheric momentum (wind) to the mass (pressure), such that the horizontal wind runs parallel to the axis of tilt of constant-pressure surfaces and the speed is proportional to the angle of the tilt; atmospheres governed by this coupling are said to be in geostrophic balance. The trick is to make horizontal temperature maps using an infrared camera, because horizontal temperature gradients mark regions where constant-pressure surfaces are not parallel, and hence mark regions where the horizontal wind speed changes with height. In Jupiter's upper troposphere, the anticyclonic structures (high-pressure regions), including both the long-lived vortices like the Great Red Spot as well as the zones themselves, are cooler than their surroundings, about 8 K cooler in the case of the Great Red Spot, whereas the cyclonic vortices and belts are correspondingly warmer than average. Both of these trends signify that the winds on Jupiter are decaying with height in the upper troposphere. For example, it is inferred that the peak cloud-top winds in the Great Red Spot, which were about 120 m/s during the Voyager encounters, drop to zero in the stratosphere at around 50 mbar. Numerical models have demonstrated that this trend helps to stabilize the winds against shear instability.

Just as in Earth's atmosphere, ultraviolet light from the Sun heats up Jupiter's upper atmosphere to form a stratosphere; the difference is that ozone is

the primary UV absorber on Earth whereas methane and a long list of hydrocarbons are the UV absorbers on Jupiter. The abundance of photochemical products in Jupiter's stratosphere makes its chemistry easier to probe spectroscopically than either the troposphere below or the thermosphere above. An entire array of hydrocarbons has been identified spectroscopically: ethane, acetylene, methylradicals, ethylene, methylacetylene, allene, propane, diacetylene, benzene—the list continues—plus stable isotopes of the above made with ^{13}C and D (deuterium). There are also trace oxygen compounds in the stratosphere, probably supplied by cometary impacts; the most prominent are carbon monoxide, carbon dioxide, and water vapor; and there are other compounds involving nitrogen, phosphorus, and sulfur. A hotbed of disequilibrium chemistry exists near each pole in the auroral regions that forms polar haze layers.

Galileo Probe. The Galileo Probe mission lasted an hour on 7 December 1995 and returned 3.5 Mbits of data to a depth of 200 km below Jupiter's cloud tops. Probes have been successfully deployed in planetary atmospheres before, including Venus, Earth, and Mars, but the success of this experiment was a milestone in spacecraft engineering because the Probe entered Jupiter's upper atmosphere at a speed of 47 km/s, the fastest atmospheric entry ever attempted. The only serious problem to occur was that the Probe's heat insulation failed to maintain the temperature inside it within the design-specification range of -20° to $+50^\circ\text{C}$ and the temperature varied at a rate of -5° to $+4^\circ\text{C}/\text{min}$, which was three times more rapid than expected; so the instrument readouts had to be recalibrated using identical twins on the ground (2).

A monitoring campaign at NASA's Infrared Telescope Facility (IRTF) in Mauna Kea, Hawaii, showed that the Probe hit the southern rim of one of the dozen or so 5- μm wavelength hot spots that drift along latitude 7°N at any given time. Hot spots are so-called because they appear bright in infrared images, signaling that they are cloud-free skylights that allow the radiation from the interior to escape to space ("hot" refers to the IR brightness temperature). The Probe's atmospheric structure instrument, cloud-particle detector (nephelometer), and other instruments show that it did not encounter the expected three-tiered layers of ammonia, ammonium hydrosulfide, and water clouds, but instead saw only the barest hint of any clouds at all. Interestingly, the locations of condensibles were pushed down into the planet in an organized fashion without being mixed together. One theory is that the hot spots are the troughs of large-amplitude waves that encircle Jupiter just north of the equator and that this wave system includes the white equatorial-plume structures.

Analyzing the Doppler shifts in the Probe's telemetry allowed determining its precise line-of-sight motions. The largest contributions came from the Probe's vertical descent and the rotation of Jupiter relative to the Galileo Orbiter overhead, but once these effects were accurately removed, a clear signal of the horizontal winds emerged. The Probe encountered an eastward (prograde) wind of 100 m/s at the cloud-top level, as expected, and then the wind speed increased rapidly to approximately 170 m/s at 4 bars; after that, it maintained a similar speed down to the 21 bar level, where telemetry was lost. Modest vertical winds of 1–2 m/s were encountered that are consistent with buoyancy waves (internal gravity waves) in the atmosphere similar to those in Earth's atmosphere.

What happened to the Galileo probe after it accomplished its 1-hour mission? Because the Probe was vented, it is thought that only a few sealed boxes inside it were crushed once high pressures were encountered, but the rest of the aluminum/titanium structure was probably able to equilibrate to the rapidly increasing pressures during descent. Thus, the ultimate fate was most likely vaporization by the intense temperatures of Jupiter's interior. First to melt was probably the parachute, made of Dacron, when the probe encountered temperatures exceeding 260°C, estimated at less than an hour after its last received transmission. Once the parachute failed, the probe's descent hastened, and before 2 hours had elapsed after the end of the telemetry, in an ambient pressure exceeding 280 bars, the Probe's aluminum components reached 660°C and melted. The melting temperature of titanium is 1,680°C, so the housing survived for about another 6 hours before melting and breaking up into droplets at about 2000 bars. The droplets then rained downward for another hour or so until they evaporated. From the time it started returning its telemetry until it was completely vaporized, the Probe probably lasted only about 10 hours, or about 1 Jupiter day (3).

Stability of the Jets. Part of the mystery of why Jupiter has so many alternating jet streams, and why they are rock steady, comes from not being able to see the abyssal circulations hidden beneath the cloud tops; this leaves open many possibilities. For example, Jupiter's jets could be seated in the visible atmosphere, with the internal circulations spun up to a greater or lesser degree as a consequence. Or they could be rooted in the convecting interior where the observable patterns in the cloud tops are shaped by conditions below. Or they could be differentially accelerated at all levels by external satellite tides. Each of these possibilities is competently represented in the scientific literature, and even a mix of all three is possible. The first possibility is that the jets are shaped in the shallow atmosphere because there are intrinsic length scales that bear on the undulation of Jupiter's constant-pressure surfaces with respect to latitude and hence, on the spacing and amplitude of its jet streams. One is called the deformation radius, $L_d = c / |f|$ (valid away from the equator), the ratio of the gravity-wave speed, c , to the Coriolis parameter, $f = 2\Omega \sin(\text{lat})$, where Ω is the planet's angular velocity. This length scale is attributed to Rossby, when used in the meteorological context and is mathematically analogous to the Debye shielding length in the theory of magnetized plasmas. For vortices, it is the distance beyond which they do not strongly feel each other's presence. The deformation radius, it is estimated, is of the order of 2000 km in Jupiter's troposphere, but probably has natural variability of the order of 50% because it is sensitive to the distribution of water vapor, which appears to be quite heterogeneous on Jupiter. Several studies suggest that it is not L_d , but rather a second length scale attributed to Rhines that sets the spacing of jets, which may be written $L_\beta = (Ua/\Omega)^{1/2}$, where U is the scale of the horizontal velocity and a is the radius of the planet. In the 1990s, it was shown numerically that persistent jets form in both a wide channel and a full spherical shell whose spacing is controlled by L_β and furthermore, that the number of jets that emerge qualitatively matches the different numbers observed for Jupiter, Saturn, Uranus, and Neptune.

But, the devil is in the details, and there is evidence to suggest that these shallow-atmosphere models all miss one key feature of Jupiter's jet streams,

namely, they all fail to reproduce the sharpness of Jupiter's westward jets. Jupiter exhibits several westward jets with curvatures that are a factor of 2 or more greater than the beta parameter, $\beta = df/dy$, where y is latitude (expressed in the local-Cartesian sense in units of length). Such sharp westward jets appear impossible to achieve in shallow-atmosphere models where there is little or no deep circulation, because the shear flow tends to become unstable through the process of barotropic shear instability. The dynamics is anisotropic because of the planetary rotation, and there not a similar restriction on the shape of shallow eastward jets. However, obtaining an equatorial jet that is eastward, as on Jupiter and Saturn, also seems to be difficult to obtain from shallow-atmosphere models, whereas the westward equatorial jets of Uranus and Neptune are not a problem.

If instead, the jets extend into Jupiter's interior, then important new possibilities arise. The Galileo Probe results show that the winds at 7°N do extend into the planet. Furthermore, an analysis of Voyager wind data in and around the Great Red Spot and White Oval BC has produced a family of abyssal circulations covering the latitude range 10°S to 40°S that have the following testable property: a model GRS placed over one of these deep-wind profiles reproduces the observed variation along streamlines of absolute vorticity, $\varsigma + f$, where ς is the relative vorticity (the vertical component of the curl of the velocity). These empirically determined circulations happen to correspond to the special case $L_d = L_\beta$, when the latter is written $L_\beta = (u/Q_y)^{1/2}$, where Q_y is the gradient of the potential vorticity (quasigeostrophic), which includes the planetary-vorticity gradient or beta parameter, $\beta = df/dy = 2(\Omega/a) \cos(\text{lat})$ as above; the relative-vorticity gradient, $d\varsigma/dy \approx -d^2u/dy^2$; and the vertical-shear (baroclinic) term, which is needed when L_d is much smaller than the planet's radius, as is the case for Jupiter. This special condition may alternatively be stated $u = Q_y L_d^2$, and it has significance for two reasons. First, it corresponds to the case of marginal shear stability with respect to a stability criterion that traces back to Kelvin, but is notably absent from most meteorology textbooks, and second, it implies significant abyssal circulations that differ from the cloud-top winds.

Cross-Disciplinary Physics. Considering that, at some level, there is perfect correspondence between the mathematics that describes the dynamics of rapidly rotating atmospheres and that which describes magnetized plasmas, the type that arises in fusion power-generation studies and have economic significance, it is particularly interesting that the same shear-stability theorem just mentioned also arises in fusion-related experiments. For example, O'Neil and Smith (4) discuss the two branches of shear stability theorems pioneered by Kelvin and proved for nonlinear (large amplitude) perturbations by Arnol'd in the mid-1960s. They point out that one branch is much easier to establish than the other and hence, is much better represented in the plasma literature, but that the harder one is needed to understand shears in their nonneutral plasma column. Completely parallel to this situation, most meteorological textbooks establish to one degree or another the stability theorems attributed to Fjørtoft, Charney, and Stern, and Rayleigh and Kuo (the barotropic stability criterion), which we have listed in order of decreasing generality and which are all related members of the easier branch, while leaving out any word about the second branch, referred to in the literature as Arnol'd's second stability theorem. It is tantalizing that when

fusion researchers need a long-lived vortex and stable shears in their plasma column, they have the option of copying how Jupiter does it.

Deep Winds. The other reason for the significance of marginal stability of Jupiter's winds to Arnol'd's second stability criterion is that this condition implies strong, alternating jet streams in Jupiter's interior that differ from those seen in the cloud tops. By assuming that the 454 m/s speed of the dark ring seen propagating outward from each of the Comet Shoemaker \simeq Levy 9 impact sites is the gravity-wave speed in Jupiter's atmosphere—not a firmly established fact—a unique member of the family of abyssal circulations mentioned before is singled out that leads to the pre-Probe prediction that Jupiter's westward jets change little with depth but that its eastward jets increase in strength by 50–100% with depth. This prediction for the eastward jets closely matches the results of the Probe's Doppler wind experiment, with the caveat that the Probe's latitude of 7°N was too close to the equator for the strong Coriolis effect assumed by this (quasigeostrophic) theory.

If Jupiter's jets are deep, then they must be stable simultaneously in the two different geometries that they span, the atmosphere and the interior. There is a promising lead regarding the atmosphere in the form of Arnol'd's second stability criterion, sharp westward jets and all. But the necessary deep winds for this would be prone to barotropic instability themselves, so it seems we have jumped out of the frying pan and into the fire. That would be the case were the deep winds subject to the thin spherical-shell geometry of the atmosphere, but they are not, as illustrated by work on the stability of deep-seated jet streams. Several groups have considered the possibility that Jupiter's jet streams are rooted deeply where the planet's internal heat source drives convection and where there is no confinement of motions inside a thin spherical shell. The problem is made all the more intriguing by the Taylor–Proudman effect, which inhibits motions in the direction parallel to the planet's rotational axis. A series of studies have shown that deep-seated convection can generate alternating jets at the top of a convecting sphere. Numerical simulations in the 1990s of a rapidly rotating, deep fluid shell achieved a broad eastward flow at the equator and alternating jets at higher latitudes. However, the jets are barely discernible through the large noise of the convection in the work to date. It now seems probable that both a deep-spherical interior and a thin, stable atmosphere coupled together are needed to model Jupiter properly. One recent coupled model has thermally driven convection in the interior that evolves to concentrate motions via “teleconvection” in the stable outer layer. More coupled atmosphere–interior experiments are needed to determine which regime ultimately picks the scale of the jets on the gas giants, or whether it is a negotiated deal.

The third class of hypotheses for Jupiter's jet streams involves the intriguing possibility that the winds are shaped and accelerated not by internal forces as before, but by satellite tides. If the interior of Jupiter on average is modestly statically stable to convection, tides can couple to it that are dominated by higher order Hough modes. These tend to produce banding that has alternating accelerations of the order of $1 \text{ cm s}^{-1} \text{ day}^{-1}$. The dominant tides come from Io, Titan, Ariel, and Triton, respectively, for Jupiter, Saturn, Uranus, and Neptune. This idea is an area of active research, and it is motivating the search for observational evidence of a tidal response at Jupiter's cloud level.

Interior

The inside of Jupiter has to be a fascinating place. Whatever the abyssal circulations are that fill the interior, they must certainly be laced into intricate convective patterns that are shaped by the planet's rapid rotation and strong magnetic field. Most of the interior is metallic hydrogen, which is conductive and the root of Jupiter's strong magnetic field. This is magnetohydrodynamics (MHD) at its best, and although we have the barest inkling of what is going on, there is great hope in uncovering more, the closer we scrutinize Jupiter.

Jupiter is an order of magnitude too small to fuse hydrogen in its core and thus be classified as a star, but it does emit more radiation than it absorbs. Only 60% of Jupiter's emitted infrared flux is reradiated solar energy; the rest percolates upward from deep inside the planet, the result of settling and the release of gravitational potential energy that continues 4.6 billion years after the formation of the planet. The temperature reaches about 10,000 K in the interior; the exact maximum value and its profile with depth are not precisely known because they are sensitive to the detailed composition. For example, if Jupiter's interior contains reduced amounts of the alkali metals, sodium and potassium, then the heat-transfer mechanism could change from convection to radiation where the temperature is in the range 1200–1500 K and the effective opacity of hydrogen and helium dips; such a radiative zone would have implications for the overlying dynamic meteorology and shear stability. But more likely, the alkali metals are not depleted, and they maintain opacity across this temperature range such that the heat-transfer mode is convection throughout Jupiter's interior.

A low-flying spacecraft that orbited Jupiter a few thousand km above its cloud tops could measure gravitational anomalies on the order of 1 mgal (10^{-5} ms^{-2}) and determine the high-order, spherical-harmonic coefficients of the gravity field, thereby illuminating the interior structure of Jupiter; such a mission is on the drawing board. Two other classes of observations that might open up the study of Jupiter's interior, in the same fundamental way that helioseismology revolutionized the study of solar interiors, are the detection of free oscillations and of tidal responses to satellites. Both are being actively pursued by observers, but the signals are weak and to date there have not been clear results. On the other hand, the entire Jovian system is affected by the makeup of Jupiter's interior, whether the influence be gravitational, magnetic, chemical, thermal, or a combination, and we can turn this fact around to infer properties of the interior indirectly.

Hydrogen: An Alkali Metal. The elements that occupy the leftmost column on the periodic table are the alkali metals, and hydrogen is a member of this group. Converting molecular hydrogen, H_2 , into monatomic metallic hydrogen in the laboratory is a modern feat that requires extremely high pressure. Most of the interior of Jupiter exceeds this pressure threshold, and hence the behavior of metallic hydrogen has important implications for our understanding of Jupiter's interior structure and the generation of its magnetic field. By briefly creating shock pressures up to 180 GPa (1.8 Mbar) and temperatures up to 4000 K in the laboratory, researchers have been able to measure the electrical conductivity of fluid metallic hydrogen. Currently, there are two differing sets of results that disagree to a level that impacts Jupiter-interior modeling, but it is likely that the

picture will clear up soon because more experiments are being performed. Present indications are that the change from molecular to metallic hydrogen is not a first-order phase transition, which would imply an abrupt boundary in Jupiter's interior and be significant for both atmospheric dynamics and chemistry, but rather is continuous and complex. Some caution must be exercised because the laser-shocks created in the laboratory may yet turn out to be supercritical. The current picture (5) is that hydrogen begins to dissociate around 40 GPa, to form significant metal-like electrical conductivity around 140 GPa; significantly it has the same value of conductivity as the fluid alkali metals Cs and Rb undergoing the same transition, and becomes completely metallic at around 300 GPa. The pressure 140 GPa (1.4 Mbar) corresponds to a depth below Jupiter's surface clouds of only about 10% of the planet's 71,492-km equatorial radius (earlier it was thought that the depth of the metallic transition was closer to 25% of the radius). Thus, approximately $(0.90)^3 = 73\%$ of the planet's volume contains metallic hydrogen. Mapping Jupiter's higher order magnetic field components using a magnetometer carried by either a satellite in low orbit or a ramjet flying in the atmosphere will provide new information about the interior structure.

Jupiter's magnetic field is similar to Earth's in many ways. The dipole tilt for Jupiter is 9.6° , which is only 2° less than the current value for Earth, and if the low-order moments for both Jupiter and Earth are extrapolated inward to their point of origin, the relative strengths of the moments are the same for both planets. Thus, a similar dipole-style dynamo process that involves convection of a conductive fluid is probably producing the magnetic fields in both planets. One major difference is that Earth's core-mantle boundary, whose position and character are precisely known from seismology, is abrupt, whereas the outer envelope of Jupiter's dynamo may not be so. Uranus and Neptune are much different; they have larger quadrupole moments, larger dipole offsets, and larger tilts, so they may have quadrupole-style dynamos. The tilt of Saturn's field is enigmatic; it is less than 1° ; the Cassini orbiter will revisit the question of why Saturn is unique in this regard.

Helium is the most important constituent in Jupiter's atmosphere after hydrogen, just as in the Sun. A primary result of the Galileo Probe was to determine accurately the helium mole fraction, which is $13.59 \pm 0.27\%$. This value is less than that inferred as the original fraction based on models of the Sun's history; so some of the helium has probably rained toward the center of Jupiter. Because neon tends to dissolve in helium drops, it is consistent that the neon mixing ratio is lower on Jupiter than in the Sun. As a rule of thumb, the Galileo Probe found that, except for oxygen, the most common elements heavier than hydrogen and helium on Jupiter are about three-times enriched compared to the Sun. The oxygen concentration on Jupiter appears to be at least solar but is complicated by the fact that it occurs as water, which is a major player in the atmosphere's dynamic meteorology, and consequently its distribution is heterogeneous.

Satellites

Starting on the inside and working out, Jupiter's largest sixteen moons are Metis, Adrastea, Amalthea, and Thebe; the four Galilean satellites Io, Europa,

Ganymede, and Callisto; then Leda, Himalia, Lysithea, Elara, Ananke, Carme, Pasipha, and Sinope. The Galileo Orbiter has vastly enhanced our understanding and appreciation of these worlds, and there is amazing diversity among the siblings. Io has the most colorful coat and the most influence on the physics of the Jovian system. But, when the topic turns to liquid water and the possibility of extraterrestrial life inside the solar system, Europa is uppermost on everyone's mind. And then there is the largest, Ganymede, which outclasses two planets and has its own magnetosphere. The smaller satellites are active too; the innermost ones are involved in shaping Jupiter's ring system. Each satellite has a different story to tell (6), but we begin with the technicolor dreamcoat.

Io. The innermost Galilean satellite has stood out for almost a century as an enigmatic moon. Laplace (*Mecanique Celeste*, Vol. 4, 1805) studied the intriguing 4:2:1 resonance among the orbital periods of Io, Europa, and Ganymede that now bear his name. In 1927, it was noted that Io (Fig. 3) has a pronounced variation in brightness with orbital phase angle. In 1964, there was a report of an anomalous brightening of Io's surface as it emerged from behind Jupiter. Also in 1964, the Australian meteorologist Bigg discovered that radio waves in the decametric wavelength range emitted by Jupiter showed a modulation that was related to Io's orbital position relative to the observer. As we now know, the interactions between Io, the plasma torus that encircles Jupiter and is supplied by Io's volcanism, and Jupiter's magnetosphere are complex and contain many feedback mechanisms. In 1971, Io occulted a bright star, providing an accurate determination of its radius and density, both of which are about 5% larger than Earth's Moon—they could be twins. However, whereas the Moon has a global heat flow of about 0.02 Wm^{-2} , Io's heat flow is 1 to 3 Wm^{-2} . At the start of the 1970s, observers were looking for evidence of a Moon-like surface, perhaps covered with some water or ammonia frost. What was found was discordant photometry at

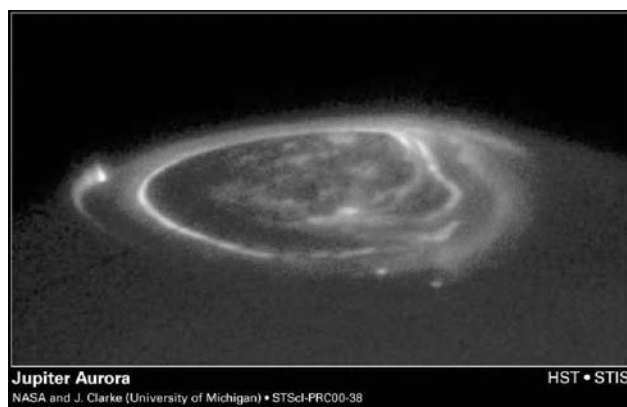


Figure 3. Hubble Space Telescope UV image of Jovian northern aurora: The polar cap. Auroral footprints can be seen in this image from Io (along the left-hand limb), Ganymede (near the center), and Europa (just below and to the right of Ganymede's auroral footprint). These emissions, produced by electric currents generated by the satellites, flow along Jupiter's magnetic field and bounce in and out of the upper atmosphere (Hubble Space Telescope/NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

visible and infrared wavelengths, sodium D line emission, and spectral evidence for sulfur, none of which is Moon-like. The Pioneer spacecraft encounters revealed that Io has an ionosphere and a thin atmosphere. In 1975, 2 years before the Voyager encounters, strong absorption near $4\text{ }\mu\text{m}$ was detected that later proved to be sulfur dioxide. In 1979, in support for Voyager, but prior to the encounters, observers discovered intense temporary brightening in the infrared from 2 to $5\text{ }\mu\text{m}$ and evidence that some of Io's surface is at 600 K compared to the daytime average of 130 K; this result was met with skepticism.

And then, just 1 week before the 5 March 1979 Voyager 1 encounter, S. Peale, P. Cassen, and R. Reynolds (7) published a paper entitled "Melting of Io by tidal dissipation," in which they predicted, "Consequences of a largely molten interior may be evident in pictures of Io's surface returned by Voyager 1," and "widespread and recurrent surface volcanism would occur" as a result of Io's role in the Laplace resonance (7). The general mechanism is as follows. Were it not for the resonant forcing of the Laplace resonance, Io's orbit would have an eccentricity of only $\varepsilon = 0.00001$, which would produce negligible tidal heating, but the value is elevated to $\varepsilon = 0.0043$. To the eye, this is still a small eccentricity, and the corresponding elliptical orbit still looks like a circle, but one that has Jupiter not quite in the middle because it is centered on one of the two foci. The main heating comes from the fact that Io turns on its axis with clockwork precision at the average rate of its orbit—it keeps the same face pointed toward Jupiter just as the Moon does to Earth—but slides faster through its orbit when closer to Jupiter (perijove) and slower when farther away (apojove), so that it nods back and forth, causing the tidal stresses and driving the most active volcanism in the solar system.

Voyager 1 found widespread volcanism on Io and no impact craters. The dramatic prediction and swift confirmation of Io's volcanism is one of the major success stories in theoretical planetary science. Today, there exists a large amount of ground-based and spacecraft observations, including more than a half-dozen close encounters of Io by the Galileo Orbiter that provide a tantalizing view of the plasma physics and magnetohydrodynamics in the Jovian system. Two major classes of volcanoes have been identified on Io. Pele-type (Fig. 4) eruptions are large, up to 300 km high, and deposit relatively dark red material; they occur in a restricted region from longitude 240° to 360° . Prometheus-type eruptions are 50–120 km high, last months or years, and deposit bright white materials; they occur all around an equatorial band.

Sulfur dioxide frost covers much of Io's surface, and volcanism and sublimation supports a tenuous SO_2 atmosphere. Beautiful UV auroras have been observed in Io's atmosphere. Some of that atmospheric mass is injected into Jupiter's magnetosphere at a rate that depends in part on the level of Io's volcanic activity and supplies electrons and S^+ , S^{2+} , O^+ , and O^{2+} ions into a torus of plasma that orbits Jupiter and is appropriately called the *Io plasma torus*. The plasma is quickly caught up in Jupiter's magnetic field and then orbits in the same 9.9-hour period at which the planet rotates; this is much faster than Io's 42.5-hour orbital period. The result is that the plasma streams past Io at a relative speed of about 60 km/s and as a consequence, generates an enormous, 10-million-ampere electrical circuit. A giant, arching structure called the *Io Flux Tube* connects Io and Jupiter along Jupiter's magnetic field lines. The location

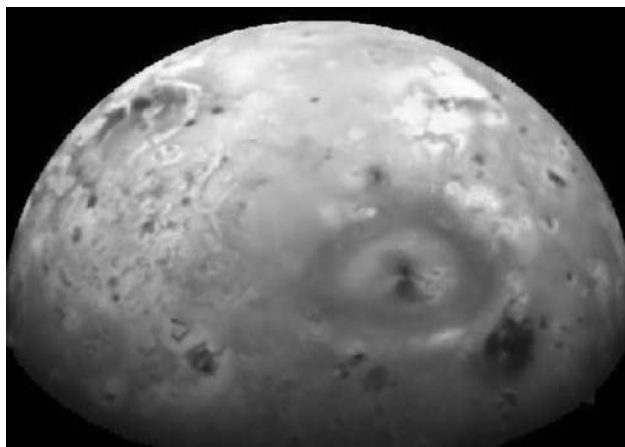


Figure 4. Galileo spacecraft image of Jupiter's moon Io. The giant active volcano Pele is prominent in this image. Much of Io's surface is covered by sulfur dioxide frost, and many of the colors may represent allotropes of sulfur (Galileo Project/NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

where this tube enters Jupiter's atmosphere near each of Jupiter's poles is called the Io footprint and is the site of intense, localized auroral emissions from Jupiter's ionosphere. The most spectacular images are observed in the infrared at the 3.4- μm wavelength, because the methane in Jupiter's atmosphere strongly absorbs, and hence eliminates, any light coming up from below the ionosphere at this wavelength, whereas the main ion present, H_3^+ , produces several strong emission lines in the same wavelength region. Io's footprint has been located in UV and visual-wavelength images as well as in IR images. The physical mechanisms involved in the interaction between the ionospheres of Io and Jupiter are not fully understood.

Ganymede. The largest planet in the solar system fittingly has the largest satellite, Ganymede. It is larger than Pluto and Mercury and maintains its own magnetosphere, like a bubble inside of Jupiter's magnetosphere; it has a field strength an order of magnitude stronger than the typical 120-nT Jovian field strength at Ganymede's orbit. There is a trace atmosphere of oxygen coming from the 50–90% ice surface. Data from two Galileo close encounters reveal that Ganymede (Fig. 5) has the lowest measured moment of inertia of any solid body in the solar system, meaning that most of its dense material has sunk to the center, probably forming a silicate core with an iron center and leaving a thick icy mantle outside. It is not surprising that Ganymede's geology is as complicated as any planet's. About 60% of the surface has been reworked into bright, grooved terrain and shows evidence of strains larger than 30% that have torn its craters apart. The other 40% of the surface is heavily cratered and is darkened by a thin layer of residue or slag that is probably left over from sublimation of the ice.

Callisto. A surprise of planetary exploration has been that each of the more than 60 moons in the solar system is different. A case in point is Callisto, which unlike its three Galilean-satellite siblings, has a heavily cratered surface and almost no trace of other geologic processes, except that some of its small craters



Figure 5. Galileo spacecraft image of Ganymede showing a line of 13 closely spaced craters. A fragmented comet similar to Comet Shoemaker–Levy 9 probably caused this feature (Galileo Project/NASA).

(<1 km) appear to have been broken apart into large blocks by an unknown agent of erosion. One theory for the uniqueness of Callisto is that it is the only Galilean satellite that has never experienced orbital resonance with its siblings. Callisto's large craters tend to be fatter than craters seen in the inner solar system. Spectroscopic observations show that there is plenty of ice worked into the surface, about 50% on average; this is also not Earth's Moon, but there are regions on Callisto that are free of ice. Not much can be said about its interior structure, to what extent it is differentiated into a silicate core and an icy mantle, but its moment of inertia is somewhat lower than that of a uniform sphere, implying some differentiation. Given Callisto's old-man appearance compared to Europa and Io, it was something of a surprise when two Galileo close encounters suggested that Callisto actively responds to the constant changes of Jupiter's magnetic field (due to Jupiter's rotation and magnetic-field tilt). An induced dipole that flips with the forcing from Jupiter is evident, which implies a conducting layer somewhere inside the satellite. A trace amount of ammonia would have a strong antifreeze effect and is a distinct possibility for keeping a layer of liquid inside Callisto and could also explain the erosion of the smallest craters. But researchers are not sure how the satellite could avoid greater differentiation if this is the case. A lander might be able to solve the mystery by producing a longer magnetic field record and by sampling the surface.

Europa. Europa is one of the smoothest spheres ever produced in nature. One has to search hard to find its handful of craters. Europa is tidally heated by the

same Laplace resonance that drives Io's volcanos, and it is clear that its surface has been worked and reworked many times. Some of the surface is mottled and composed of a jumble of blocks, called chaos terrain, that probably indicates local melting of ice. On the large scale, there is a remarkable series of dark lines that crisscross the surface. Many causes may contribute to the stress field that produces these, including nonsynchronous rotation, polar wander (a tendency for the polar ice to thicken such that the ice shell's global orientation becomes unstable), global contraction or expansion, or tidal stress. Europa gets between 100 and 1000 times the radiative flux that Earth's Moon receives from the solar wind, causing significant chemical changes in new surface material in less than a decade.

The magnetic signature around Europa suggests that there is an induced dipole from Jupiter's time-variable magnetic field, which implies a briny sub-surface ocean. But the data are enigmatic, and a fixed dipole cannot be ruled out. Taken as an ensemble, essentially all of the features associated with Europa's geology, magnetic response, and even the presence of salty contaminants on its surface, suggest that liquid water lies beneath the surface. However, each indicator can also be explained individually without liquid water, so one must guard against overinterpretation. Even so, the possibility of an extraterrestrial liquid ocean has policy implications; for example, after its useful lifetime, the Galileo orbiter will be scuttled into Jupiter's atmosphere to eliminate any possibility that it might one day collide with Europa. A new mission to Europa is high on the NASA's priority list; the goal is to settle definitively the question of a liquid ocean. Part of the excitement lies in the fact that ecosystems exist on Earth that do not depend on sunlight; the most relevant are the varied life-forms that derive their energy from "black smoker" hydrothermal vents in the deep oceans. It is not hard to conceive of such vents inside of Europa, but there is a problem involving the chemistry needed to harness that energy. The oxidants in Earth's hydrothermal vents are ultimately tied to atmospheric oxygen through the overturning action of plate tectonics. Without such cultivation, Europa's mantle would become a reducer rather than an oxidizer, and its vents would emit methane (CH_4) rather than carbon dioxide (CO_2), which would be poisonous to Earth's vent ecosystems and presumably also to Europa's. But, when the subject is life, at least on Earth, there always seems to be a way; for example, rare bacteria exist that use hydrogen to reduce minerals, and similar organisms might not find Europa completely inhospitable.

Small Satellites. Jupiter plays host to at least an additional 30 small, outer satellites and also to Trojan asteroids that lead or trail the planet by 60° in the same orbit (at the stable L4 and L5 Lagrange points, respectively). The orbits of the outer satellites tend to be large, have high eccentricity and high inclination, and often have retrograde orbits. The gravitational reach of Jupiter, its Hill sphere, is about one-third of an AU or just over 700 Jovian radii. The orbits of the outer satellites average about half this distance, which makes them prone to perturbations. Not much is known yet about their chemical composition, but there is the suggestion in color data that they may form groups of fragments that come from distinct parent bodies. Jupiter also has a tendency to grab comets into an orbit around itself at the rate of a few per century. The most famous is Comet Shoemaker-Levy 9, which was captured by Jupiter and immediately afterwards

collided with the planet; every piece of telescopic glass in the solar system was pointed to witness the event. Such collisions may occur a few times per millennium. Jupiter's Trojans could possibly be snow-line planetesimals, in which case they would be as valuable as comets for uncovering facts about the protoplanetary nebula.

Rings and Dust. One goal of the Voyager 1 encounter with Jupiter was to discover the existence of any faint ring or rings (Fig. 6). No major rings had ever been seen in the backscattered sunlight that is visible from Earth's inferior orbit, but this did not rule out a ring made of dust, which is best viewed in forward-scattered sunlight, as anyone will attest who has driven into a sunrise or sunset with a dusty windshield. The goal was met handsomely by a single image of the dark side of Jupiter where the forward-scattered sunlight illuminated a distinct ring. The same method was successfully used by Voyager 2 at Uranus and Neptune to capture images of the dusty components of their narrow-ring systems. The discovery at Jupiter was followed up 4 months later by an imaging sequence, customized on-the-fly for Voyager 2. Today, Jupiter's ring can be imaged directly from Earth using infrared telescopes.

All four gas giants have ring systems, whereas none of the terrestrial-class planets have them; this suggests that a ready supply of orbiting mass is needed to make a ring. Saturn's ring system is the most majestic in the solar system, but each shows structure as a function of orbital radius, sometimes gaps and sometimes regions of enhanced density, that are the result of resonances between the orbital elements of the ring particles and a particular satellite, or pair of satellites. Jupiter's dusty ring system harbors an additional feature; it contains resonances with Jupiter's inclined and rotating magnetic field, called Lorentz resonances, a prime example of the way the dynamics of orbiting dust is affected by electromagnetic and gravitational forces simultaneously. There are three

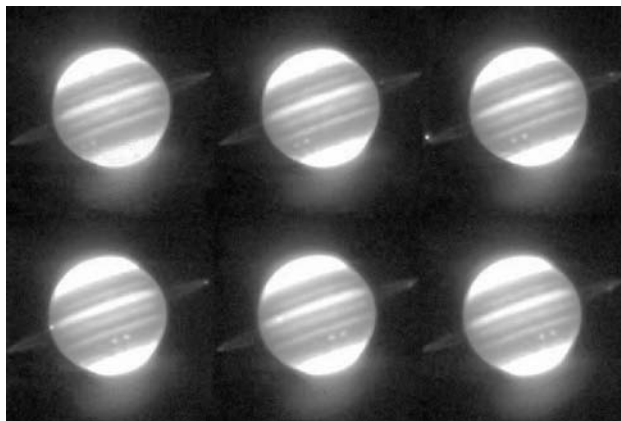


Figure 6. Sequence of Infrared Telescope Facility (IRTF) images of Jupiter, its ring, and the two inner satellites Metis and Amalthea. The images span 2 hours; time increases from upper left to lower right. Metis first appears in the second image, following transit across the face of the planet. The brighter satellite, Amalthea, first appears in the third image before transiting across the planet (Infrared Telescope Facility/NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

components of Jupiter's ring system. The main ring is flat and has a sharp outer edge at an orbital radius of 129–130 km and is about 7000 km wide. The ring's dust supply comes from Adrastea and Metis. The two other satellites orbiting outside the main ring, Amalthea and Thebe, guard two faint rings called the gossamer rings; Galileo images reveal structure even in these rings. The out-of-plane thickness of each ring matches the inclination of its associated satellite, and Galileo images show dust coming from Amalthea and Thebe that supplies the gossamer rings. The third component is called the halo, which starts at the main ring and flares out as it extends about 20,000 km inward to the midway point with the planet's cloud tops.

Jupiter's dust environment is as well sampled as Earth's—five dust detectors have flown through the Jupiter system to date. The Pioneer 10 and 11 spacecraft flew by Jupiter in 1973 and 1974, respectively, and recorded a thousandfold increase in dust as they entered the Jovian system. Neither Voyager 1 or 2 carried a dust detector, but in 1992, the Ulysses spacecraft passed through the Jupiter system to achieve its primary mission of a high-solar latitude orbit around the Sun and carried a dust detector that had greatly increased sensitivity compared to the Pioneer instruments. It discovered bursts of dust coming from the Jovian system that extend far into interplanetary space, and these were also recorded by a similar instrument carried into Jupiter orbit by the Galileo spacecraft. Time series of the dust hits on the various detectors show a strong 5-hour oscillation that is modulated by the passing of Jupiter's magnetic equator over the spacecraft. Analysis of the dust streams emanating from Jupiter show that charged dust particles interact with the interplanetary magnetic field, so that the paths of smaller particles are bent more than the paths of larger ones, just as in a mass spectrometer.

Each time the Galileo Orbiter passed close by one of the major satellites, it detected a jump in the dust concentration, and when this happened, the speed of the dust hitting the detector matched the speed of the spacecraft relative to each satellite. This is how the existence of clouds of dust surrounding each moon was discovered. These clouds are probably stirred up by a continuous bombardment of small meteoroids. The active volcanism on Io provides a much greater source of dust, and in addition, the Io plasma torus is a major source of charge for the dust caught up in the Jovian system. Because it is an orbiter instead of a flyby spacecraft, Galileo has been able to spend years mapping in detail the dust distribution inside the Jovian system, including clouds of dust around the Galilean satellites and the interaction of interplanetary dust with Jupiter's powerful magnetosphere. A unique opportunity came when the Cassini spacecraft, bound for Saturn but making a close encounter with Jupiter in 2000 to gain a gravity assist, carried the most sensitive dust detector yet through the system. The result was that one particular dust stream was observed *in situ* at two different times and positions by Cassini and Galileo.

We can expect even closer scrutiny of Jupiter, its satellites, and its magnetosphere in the decades to come, as Europa is fully investigated and as the high-order spherical harmonics of the planet's gravitational and magnetic fields are used to peer into its interior. One day, we may hope to appreciate Jupiter's role fully as protector of the inner solar system and king of the planets.

BIBLIOGRAPHY

1. Bagenal, F., T. Dowling, and W. McKinnon (eds). *Jupiter: The Planet, Satellites, and Magnetosphere*. Cambridge University Press, Cambridge, 2003, in preparation.
2. Seiff, A., D.B. Kirk, T.C.D. Knight, R.E. Young, J.D. Mihalov, L.A. Young, F.S. Milos, G. Schubert, R.C. Blanchard, and D. Atkinson. Thermal structure of Jupiter's atmosphere near the edge of a 5- μm hot spot in the north equatorial belt. *JGR* 103: 22,857–22,889 (1998).
3. Lunine, J., and R. Young. Fate of the Galileo Probe. *Planetary Rep.* 15/6: 14 (1995).
4. O'Neil, T.M., and R.A. Smith. Stability theorem for off-axis states of a non-neutral plasma column. *Phys. Fluids B* 4: 2720–2728 (1992).
5. Nellis, W.J. Metallization of fluid hydrogen at 140 GPa (1.4 Mbar): Implications for Jupiter. *Planetary and Space Sci.* 48: 671–677 (2000).
6. Showman, A.P., and R. Malholtra. The Galilean satellites. *Science* 286: 77–84 (1999).
7. Peale, S.J., P. Cassen, and R.T. Reynolds. Melting of Io by tidal dissipation. *Science* 203: 892–894 (1979).

TIMOTHY E. DOWLING
Comparative Planetology Laboratory
University of Louisville
Louisville, Kentucky

L

LIQUID-FUELED ROCKETS

Since the beginning of the space age, liquid-fueled rockets have provided the basis of everything we have accomplished in space. Whether it has been the mighty F-1, which powered Apollo to the Moon, or the workhorse RL-10, or the high performance SSME, all of the major missions depend on liquid-fueled propulsion. And as we move forward to the next generation systems, the success or failure of these systems will be intimately tied to the success of its liquid-fuel propulsion system.

The following discussion focuses on the design of liquid-fueled rocket systems and examines the design choices that must be considered in integrating the propulsion system for a selected application. It is to be understood that liquid-fueled systems refer to the oxidizer as well as the fuel and that both of these are carried aboard the vehicle. Additional detailed information on the design of liquid propellant rocket engines can be found in References 1–3.

Today's liquid-fueled rocket propulsion designs originate from the theories of a Russian, Konstantin E. Tsiolkovsky, and from experiments performed by an American, Robert H. Goddard and a German, Herman Oberth. Goddard and Oberth made some of the first working liquid propellant rocket engines. Goddard was first to successfully test a liquid-fueled engine, which he did in 1926. Because of lack of interest by the United States government, Goddard worked primarily as a loner with a small cadre of technicians to support his work in the White Sands, New Mexico desert. He contributed to U.S. rocket development during World War II, but the importance of his contribution to rocket development was not recognized until the beginning of the space age. Herman Oberth, working with young Werner von Braun, developed liquid propellant rockets that got the attention of the German military. Their efforts culminated in the development of the military A-4 (or V-2) rocket. After World War II, Werner von Braun and many expatriated German scientists came to the United States. Their expertise and technical capabilities led to the development of the Jupiter and

Redstone missiles and the Saturn series of manned launch vehicles for the Apollo Program.

Rocket Propulsion Systems

A liquid-fueled rocket propulsion system consists of a number of carefully integrated components that must be carefully chosen and designed to function as part of the whole. The operation of an integrated system can best be viewed by identifying the required functions that must be satisfied and then considering the components that are chosen to fulfill these functions. The primary function of a rocket propulsion system is to produce thrust that accelerates the spacecraft to the final velocity required to achieve its goal. During the flight, the thrust must be controlled in magnitude and direction, allowing the space vehicle to be accelerated and guided along a desired flight trajectory. To produce thrust, the propellants are burned in a combustion chamber at high pressure, yielding high temperature gases that are subsequently expanded and accelerated through a convergent-divergent nozzle and ejected at high supersonic velocity. Thrust is produced in the direction opposite to the mean flow of the ejected hot gas at a magnitude equal to the difference in momentum between the incoming liquid propellants and the ejected hot gas plus the difference between the product of the pressure and the area at the nozzle exit and chamber.

To produce a practical propulsion system, the integrated system must control the flow of propellants from the storage tanks, deliver them at the required pressures, and inject them into the combustor. This process can be achieved either by pumping the propellant to the required pressures (pump-fed system) or by pressurizing them with high-pressure gas in the storage tanks (pressurized system). The integrated pumped system must provide the power required to drive the pumps and provide cooling for the thrust chamber (i.e., combustor chamber and nozzle) surfaces exposed to the hot combustion gases. For pressure-fed systems, a gas pressurization system must be provided. The nature of the component requirements and the relevant design options depend on the application for which the system is being designed. The design process for the propulsion system must consider the following questions:

- What is the spacecraft application?
- What are the vehicle's operating requirements?
- Which fuel and oxidizer should be used?
- Which type of propellant supply system?
- What type of propellant pressurization?
- What provides the engine pumping power?
- Which type of thrust chamber design?

Part of this article is structured around the subsystems of an engine. But it is important to keep in mind that these engines are highly coupled systems, from both the viewpoint of mechanical systems and fluid flow. For instance, in a pre-burner cycle, the size of the pumps defines the fraction of the propellants used to

drive the turbopumps, and this combined with the selected preburner propellant mixture ratio defines the maximum combustion chamber pressure. The preburner mixture ratio defines the turbine operating temperature and the potential need for high temperature materials, or in maximum performance designs, the need for cooled turbine components. Thus, no design of a subsystem is really complete until its interactions with the rest of the system are understood.

An overview of some of the design issues would be helpful to put things in perspective. Why consider the spacecraft application during engine design? Because requirements arising from the application become a strong driver for defining the selection of propellants and the design features of the integrated propulsion system. Should the system be designed to perform best in the atmosphere or in vacuum conditions? Are refrigerated cryogenic propellants acceptable, or is long-term storage important in the application? What level of engine combustion pressure should be considered, and what range of throttle control is important? Power to drive the pumps must be generated by the engine cycle. Which methods are available, and what is the best choice for a selected application? What issues arise from the fuel tanks and the associated fuel delivery system? Does the engine run only at launch time, or will it be used again after coasting in space, and how does this issue affect the design choices? What about the design of the combustion chamber and thrust nozzle (thrust chamber)? Which design features should be considered to yield the required engine performance and to provide adequate cooling of the structure? What considerations should be made with regard to manufacturing capabilities during the design of the engine? These issues are addressed in the sections that follow.

Definition of Designs. The following design parameters have determined the selection of specific liquid rocket propulsion on past and present launch vehicles and spacecraft:

1. Propellant chemistry. Propellant chemistry is selected to provide the highest performance for a feasible mechanical design that is safe to operate and is consistent with storage requirements.
2. Vehicle performance. Vehicle performance is maximized by optimizing the propulsion design to yield the maximum performance at the lowest weight and with the smallest geometric package to interface with the vehicle. When maximizing vehicle performance, optimization of the propellant combination affects the vehicle in two ways: (1) the propellant bulk density that affects vehicle volume and (2) the specific impulse (I_{sp}) that affects vehicle gross weight. Bulk density is the ratio of the total mass of propellants burned to their total stored volume. I_{sp} is the thrust produced per unit mass of propellants burned. To produce a specific amount of vehicle acceleration, higher I_{sp} reduces the mass of propellants required, and higher bulk density reduces the size of the propellant storage tanks.
3. Operating environment. Three major application categories exist in designing liquid rockets: boost from sea level, boost in near-vacuum or vacuum conditions, and in-orbit operations. The last category may be further divided into two subcategories, orbital maneuvering and station keeping.

4. Cost versus performance. Each engine design represents a compromise between high performance and affordability. The mission requirements and the financial realities for the planned vehicle must be considered when selecting a system design.
5. Payload. The question of essential importance here is whether or not the vehicle will be carrying passengers. It will be shown that the requirements of human-rating flight hardware have driven engine choices and affected performance, cost, and reliability.
6. Reusability of the rocket stages. Whether the hardware is to be used again or is part of an expendable system affects the choice of engine type and also the complexity of the engines.

Items 3 through 6 will be discussed immediately followed by a more in-depth discussion of 1 and 2.

Operating Environment. The requirements for launch-vehicle boost from sea level have traditionally driven first-stage systems toward high-thrust, moderate-performance engines. Kerosene has traditionally been the most common fuel for first-stage propulsion systems, and liquid oxygen is the oxidizer. The primary advantages of this choice are ease of propellant handling because kerosene is storable at ambient temperatures, and high propellant bulk density. Sea level specific impulse I_{sp} (see Eq. 5) from 250–300 seconds is typical of such engines. This is not as high as it would be with an oxygen–hydrogen system, but performance is not as crucial for first stages as it is for upper stages. Weight increase in the first stage does not necessarily force resizing the other stages. In addition, the higher propellant bulk density allows the construction of smaller fuel tanks. Because the first stage is typically the largest for any launch vehicle, decreasing its size reduces the cross-sectional area of the whole vehicle and reduces drag losses. An example of a kerosene-burning first stage is the SI-C on the Apollo Saturn V vehicle. Five Rocketdyne F-1 engines operating on a gas generator cycle comprised the propulsion for this stage. Nearly every launcher in the U.S. fleet has used Earth-storable fuel in its first stage. The first stage of the Atlas and Delta rockets are kerosene-powered; the Titan rockets use Aerozine-50 (50–50 hydrazine/UDMH) and nitrogen tetroxide as the oxidizer. The Space Shuttle Main Engine (SSME) is the major exception to this. Liquid hydrogen and liquid oxygen (LOX) were chosen as the propellant combination because of the desire to maximize performance to achieve orbit. Note that all four of the rocket families mentioned that have first stage hydrocarbon engines used cryogenic-fueled upper stages.

A large number of upper stages used liquid hydrogen (LH_2) and liquid oxygen as propellants. Cryogenic fuels are used in upper stages primarily because of the high specific impulse, which minimizes the overall stage weight needed to create the required change in velocity. Reducing the weight for a vehicle's second or third stages is highly desirable because the size of the stages below may then be decreased as well. Pioneers such as Tsiolkovsky recognized the utility of liquid hydrogen as a rocket fuel. The development of cryogenic technology accelerated after World War II, pushed by the desire to take advantage of hydrogen's specific impulse potential. The first liquid-hydrogen-fueled

rocket engine to be used on a launch vehicle was the Pratt & Whitney RL10, which debuted on the Atlas-Centaur stage in 1963 (4). The Centaur was also used on the Titan booster, and a new version of the RL10 powers the upper stage of the new Delta III. The Saturn V used Rocketdyne's hydrogen-fueled J-2 engines for its second and third stages and demonstrated the benefit of cryogenic upper stages on a vehicle that has a storable first stage.

As noted before, the third category of operating conditions is that for in-orbit operations. The need for high reliability and restart capability in propulsion dictates that the engine be as simple as possible. Typically, this requirement has been met with hypergolic propellants. These are fuels that ignite on contact and therefore do not need a separate ignition system. The specific impulse of hypergolic propellants is not as high as that afforded by hydrogen, but high performance is not crucial for in-orbit propulsion because the change in velocity is usually small. Most propellants used for in-orbit propulsion systems are also storable, a quality made necessary because they may remain in the vehicle's tanks for an extended period of time before use. A prime example of the use of storable hypergolic propellants for orbital maneuvering is the Space Shuttle's Orbital Maneuvering System (OMS). The OMS is used to provide the final velocity change necessary to inject the shuttle into its orbit, perform orbital changes during the mission, and provide the burn to reduce orbital velocity for deorbiting. These two engines use monomethylhydrazine and nitrogen tetroxide and develop 6000 pounds of thrust each. In addition to the OMS engines, the Shuttle has a reaction control system that is used to orient the vehicle in space, provide some deorbit velocity change, and separate the Orbiter from the external propellant tank. These two motors are also powered by monomethylhydrazine and nitrogen tetroxide and are pressure-fed, rather than pump-fed, which greatly contributes to their reliability and simplicity (5). Pressure-fed engines were common in boosters in early rockets flown by the Germans and by pioneers such as Robert Goddard, when launchers had only a fraction of the velocity required to achieve orbit. Pressure-fed engines are used for nearly every in-orbit application today because they are optimized for velocity requirements that are not as great as booster stages.

Cost versus Performance. The goal of any rocket development program is to achieve the highest possible performance and reliability within the financial constraints of the program. At no time have the imperatives of cost containment been so important as they are today, a fact reflected in most new engine designs. Total "cost of ownership" can be defined as the sum of the development cost, cost of procurement of production units, operational cost, and the maintenance cost. The importance of each of these various elements depends on the application requirements and on the operational goals. The cost of development is directly related to the complexity of the system. It can be reduced if systems can be developed as individual components and the developed components are subsequently integrated. In general, engines that have simple power cycles and lower chamber pressure cost less than high-pressure, high-performance systems. Interpreting this statement with regard to the type of engine cycle, gas generator cycles have a lower development cost than preburner cycles because operating pressures can be lower and power cycles can be developed independently of the thrust chamber operation.

The Rocketdyne RS-68, a liquid hydrogen/liquid oxygen engine to be used on the first stage of the Delta IV launch vehicle, is a gas generator cycle engine designed for simplicity, low cost, and moderate performance. The selection of the gas generator cycle is an important part of this cost reduction because the increase in engine simplicity and decrease in chamber pressure typically associated with gas generators reduce the cost of the system.

Alternatively, the expander cycle, used currently on the RL10 engine with liquid oxygen and liquid hydrogen, provides a means to avoid the performance losses inherent in gas generator engines, while providing potential for reduced cost. This cycle has the attributes for very low cost of manufacture and still delivers greater performance than the gas generator cycle.

On the other hand, the Space Shuttle Main Engine was designed to meet much more stringent performance requirements, without being as limited by financial restrictions as modern expendable launch vehicles. The performance requirements dictated the selection of liquid hydrogen as the fuel and staged combustion as the engine cycle. Both choices resulted in greater complexity and cost for the engine but resulted in a vacuum specific impulse in excess of 450 seconds.

It is recognized more and more that the most significant factor in cost is the design process itself. As a result, many new techniques are being used and are achieving great success. The most successful is the use of integrated product teams to execute the design and development process. These are colocated multidisciplinary teams that include every function from design engineers to suppliers. These teams ensure that producibility and inspectability are built into the product when it is being designed for performance, rather than being added later, as is much more costly and was typical of the past. Another example is using advanced computer-aided design models that allow instantaneous sharing of information by all members of the design team.

Finally, when the design team focuses on cost, it leads to decisions that may not be made otherwise. Choices that lead to a more moderate environment result in lower pressures, temperatures, pump speeds, vibrations, and heat loads. All of these have a significant impact in reducing cost.

Payload. Most rocket engines are designed to carry unmanned payloads into space. Although reliability is important in these power plants, it becomes much more crucial for human spaceflight. When lives depend on the hardware, choices are made to maximize the probability that the engine will work correctly. Reliability was the foremost concern of designers of the Apollo propulsion systems. Examples are seen in the Apollo Service Propulsion System and the descent and ascent engines for the Lunar Module. If one of the elements had failed to work, astronauts might have been stranded in orbit around the Moon or on its surface. The first choice made to increase the dependability of the Apollo systems was to make all engines pressure-fed. The absence of pumps eliminated a source of potential failures. The choice of propellants also reflected the desire for reliability. The Service Propulsion System, the engine providing the velocity change necessary to enter and leave orbit around the Moon, used nitrogen tetroxide and a 50/50 mix of hydrazine and unsymmetrical dimethylhydrazine as fuels. This combination is hypergolic, and both propellants are storable. Both engines for the lunar module also used storable, hypergolic propellants (6). Performance

requirements for the SSME dictated the propellant choice and cycle choice. Reliability was achieved through rigorous engineering development.

The choices of cycles and propellants are not the only criteria that affect performance, cost, and reliability. The selection of component materials can affect the overall engine design characteristics. One example is the nozzle on the lunar module ascent engine. The choice for this nozzle was to use ablative material that would be thermally eroded as the engine operated. This simple nozzle design that eliminated the need for cooling passages and complicated plumbing was appropriate for an engine which would have to start only once, and yet would have to operate as designed since the first time it was used a long way from Earth.

Reusability. It is widely held that the only real means to achieving the low-cost access to space that the country needs is through extensive use of reusable systems. No matter how low the cost of vehicles and engines is driven, these are still highly complex systems and will never be cheap enough to throw away after each flight.

The only current reusable system is the Space Shuttle, although this has not come close to achieving its design goal for reusability. The design specification during the original design studies for the SSME was 100 flights. Important advances in technology have occurred since the SSME was designed. There are vastly improved tools for understanding internal hot gas flow. These are needed to determine the thermal and mechanical loads (both steady and unsteady) and to predict when parts would ultimately fail. There is improved understanding of both low-cycle and high-cycle fatigue. In addition, there are advanced materials that resist these hostile environments better and thus aid in extending service life. Finally, there are new component designs, such as hydrostatic bearings. Future systems will employ many of these technologies whatever the design requirement, but reusable systems will benefit most.

As the increased emphasis on cost drives changes in design philosophy, arguably the greatest benefit to reusability will come from design decisions. Decisions that focus on life-cycle cost rather than only on development cost will lead to different systems. Similar trade-offs must be made with respect to performance. It is extremely unlikely that a highly reusable system will also provide the ultimate in performance. The challenge for today's propulsion system designer is to provide sufficient performance to achieve the mission while ensuring enough margin so that the system is truly reusable.

Propellant Choices

Propellants are typically chosen on the basis of several different parameters. In general, these are the density and the molecular weight, the storability (Earth or space ambient temperatures considered), and the ratio of oxidizer to fuel that optimizes the given combination for maximum performance and bulk density. The performance for given propellant combinations is a function of the oxidizer-to-fuel mixture ratio that maximizes the temperature of combustion, which generally maximizes the performance (specific impulse). The bulk density is the relative density of the mixture when considered as a total mass and total volume within the designed vehicle tankage. Mission requirements dictate which of

these parameters is most important. Propellant choice is often a matter of compromise between the desire for high performance and the need for ease in propellant handling and low cost. A summary of propellant properties, and some of their vehicle applications, is shown in Table 1.

Propellant Performance Drivers. Thrust is produced by using the energy released in the combustion process to accelerate the resulting gases as they flow through the thrust chamber, thereby increasing their momentum. The gross thrust F_g produced by a rocket engine is the difference in momentum between the propellants entering the combustion chamber and the products of combustion exiting the nozzle. Because the mass flow rate w_p is constant through the thrust chamber, the gross thrust is directly proportional to the difference in propellant velocities between the chamber inlet V_n and the nozzle exit V_e :

$$F_g = w_p(V_e - V_n). \quad (1)$$

The theoretical maximum exit velocity V_e can be calculated from the following equation:

$$V_e = \sqrt{2 \frac{k}{k-1} \frac{R_u}{M} T_c \left[1 - \left(\frac{P_e}{P_c} \right)^{\frac{k-1}{k}} \right]}, \quad (2)$$

where

V_e = exit velocity

k = ratio of specific heats, C_p/C_v

T_c = combustion temperature

P_c = chamber total pressure

R_u = universal gas constant

M = molecular weight

P_e = exit pressure

All properties used in this equation are those of the combustion products. For expansion to vacuum conditions ($P_e = 0$), the equation simplifies to

$$V_e = \sqrt{2 \frac{k}{k-1} \frac{R_u}{M} T_c}. \quad (3)$$

When operating in the atmosphere, the net thrust F_n produced by the engine is reduced by atmospheric pressure acting on the external surface of the engine. Thrust is reduced because the cumulative pressure acting on the outer surface of the nozzle acts in a direction opposite to the direction of gross thrust. Thrust is further affected by the nozzle exit pressure through its influence on exit velocity, as seen in Equation 2. Therefore, net thrust is calculated as

$$F_n = w_p(V_e - V_n) + A_e(P_e - P_a), \quad (4)$$

Table 1. Propellant Property References

Oxidizers	Chemical formula	Molecular Wt., kg/m ³	T _{boil} , K	T _{freeze} , K	P _{vapor} Pa	Storability	Vehicle application(s)
Liquid oxygen	O ₂	32.00	90.0	54.4	5,200 at 88.7 K	Cryogenic	Widely used for sea level or upper stage boost
Hydrogen peroxide	H ₂ O ₂	34.016	419	267.4	345 at 298 K	Earth-storable	In-orbit (extended storage)
Fluorine	F ₂	38.00	85.02	53.54	6,500 at 66.5 K	Cryogenic	In-orbit; applications limited due to toxicity
Nitrogen tetroxide	N ₂ O ₄	92.016	294.3	261.95	95,800 at 293 K	Earth-storable	Sea level boost (Titan IV: LR87) Upper Stage (Titan IV: LR91) In-orbit operations (Shuttle: OMS)
Fuels	Chemical formula	Molecular Wt., kg/m ³	T _{boil} , K	T _{freeze} , K	P _{vapor} Pa	Storability	Vehicle application(s)
Liquid hydrogen	H ₂	2.016	20.4	14.0	202,600 at 23 K	Cryogenic	Boost (Shuttle: SSME) upper stage (Centaur: RL10)
RP-1 (kerosene)	CH _{1.97}	175	460–540	225	2,275 at 344 K	Earth-storable	Boost, primarily sea level operation (Delta: RS-27, Saturn V: F-1)
Hydrazine	N ₂ H ₄	32.05	386.66	274.69	19,300 at 344 K	Earth-storable	In-orbit operations (Apollo Service Propulsion System)
Monomethyl hydrazine (MMH)	CH ₃ NH ₂ NH	46.072	360.6	220.7	60,657 at 344 K	Earth-storable	In-orbit (Shuttle: OMS)
Unsymmetrical dimethyl hydrazine (UDMH)	(CH ₃) ₂ N-NH ₂	60.10	336	216	1.213 × 10 ⁶ at 344 K	Earth-storable	In-orbit boost (Ariane 4: Viking V)
Methane	CH ₄	16.03	111.6	90.5	33,000 at 100 K	Cryogenic	Boost: sea level or upper stage
Propane	C ₃ H ₈	36.58	231.0	83.4	896,300 at 298 K	Cryogenic	Boost: sea level or upper stage
92.5% ethyl alcohol	C ₂ H ₅ OH	41.25	351	150	89,600 at 344 K	Earth-storable	In orbit

where A_e is the cross-sectional area of the nozzle exit, P_e is the static pressure of the combustion gases at the nozzle exit, and P_a is the atmospheric pressure outside the engine.

Specific impulse, defined as the net thrust produced per unit of mass flow rate, is calculated from

$$I_{sp} = F_n / w_p = (V_e - V_n) + A_e(P_e - P_a) / w_p. \quad (5)$$

At high altitudes and in space, the atmospheric pressure term goes to zero, and the specific impulse becomes the “vacuum impulse” (I_{vac}).

The unburned propellants enter the combustor in a relatively cold, dense state, so the inlet velocity is very low relative to the exit velocity, which is at high supersonic speed. Considering this difference, the relative features that affect propulsion performance, or I_{sp} , can be examined by considering the factors governing exit velocity. Examining Equation 2, the two factors that can vary significantly and have the most influence on exit velocity are total combustion temperature and molecular weight. These are primarily functions of the fuel and oxidizer chemistry and the mixture ratio. Combustion temperature only weakly depends on pressure. As shown in Equation 2, exit velocity (and I_{sp}) increases with higher combustion temperature and lower molecular weight. For a given ratio of fuel and oxidizer, there will be a resulting set of equilibrium products at a defined temperature. Equilibrium conditions are usually satisfied in conventional liquid-fueled rocket engines. Typical propellants in use today are composed of various combinations of hydrogen, carbon, oxygen, and nitrogen. The products of various propellant combinations are typically composed of mixtures of hydrogen, water, carbon dioxide, nitrogen, and oxides of nitrogen. Hydrogen and water vapor are the lightest products and produce the best performance. Therefore, fuels that are very high in hydrogen content produce better performance when burned with oxygen than hydrocarbons (i.e., kerosene) or hydrazines. Equally important, such fuel-oxidizer combinations typically yield higher combustion temperatures which also results in higher specific impulse.

Cryogenics. Liquid hydrogen is the highest performing fuel commonly used today. Paired with liquid oxygen, it can achieve specific impulses greater than 470 seconds in vacuum conditions. Liquid hydrogen can achieve such high performance due to the combination of the low molecular weight of its combustion products and high flame temperature, particularly when burned at a near optimum mixture ratio.

Maximum combustion temperature is achieved when the propellants are combined in stoichiometric proportions. For hydrogen and oxygen, this translates to a ratio of 8:1. For maximum performance, however, oxygen and hydrogen are not burned in stoichiometric proportions because that would excessively increase the molecular weight of the exhaust gases. It would also result in a chamber temperature so high that many serious hardware difficulties would be faced. For most practical applications, hydrogen and oxygen are combined in a ratio between 5 and 6:1, the range that provides the optimum combination of low molecular weight and high but acceptable combustion temperature.

Liquid hydrogen is used in applications that require high performance, such as the Space Shuttle and the Centaur upper stage, but it does possess

disadvantages. The density of liquid hydrogen is extremely low, about 4.4 lb/ft³ at saturation conditions. This value may be compared with that of another common fuel, kerosene, which has a density of about 50 lb/ft³ (7). The bulk density of a liquid oxygen–kerosene fueled system operating at a typical ratio of 2.72:1 is about 2.8 times that of a comparable liquid oxygen–liquid hydrogen fueled stage operating at a ratio of 6:1. The specific impulse achievable with kerosene is about 25% less than that for hydrogen at typical launch conditions. This results in a propellant mass requirement that is 25% less for hydrogen but is not sufficient to compensate for the large difference in bulk density between the two fuels. A liquid hydrogen rocket, therefore, requires much more tankage volume than a comparable kerosene-fueled system. For a first-stage propulsion system, the large size of hydrogen tanks may make the overall diameter of the stage too large, creating excessive drag losses, or may present significant structural challenges. For this reason, many boosters such as the Saturn V, Atlas and Titan noted before, have used hydrogen-fueled upper stages, for which the propellant load is comparatively small, and noncryogenic first stages.

Two other cryogenic fuels that have been studied for rocket propulsion applications are methane (CH₄) and propane (C₃H₈), both hydrocarbon compounds. Although both propellants are classified as cryogenic, the boiling point of propane is nearly high enough to permit that fuel to be space-storable. It is gaseous at room temperature and ambient pressure but liquid in the colder conditions possible in space. LOX/CH₄-fueled engines, in particular, have received attention as possible propulsion systems for Mars missions due to the possibility of using hydrogen and Martian carbon dioxide to form methane. LOX/CH₄ engines operating at a mixture ratio of 3.5:1 produce an I_{sp} of about 390 seconds under vacuum conditions; LOX/C₃H₈ engines, by comparison, yield a vacuum I_{sp} = 386 s with a mixture ratio of 3.2:1 (8). No propane- or methane-powered engines have yet flown on a launch vehicle or spacecraft.

Kerosene (RP-1). Kerosene fuel, otherwise known as RP-1, affords lower performance than liquid hydrogen but offers greatly increased bulk density and ease of handling. Kerosene is Earth-storable, which means that it is in the liquid state under ambient conditions at sea level. Unfortunately, the average molecular weight of the combustion products of kerosene is much greater than that of hydrogen which makes the optimum specific impulse available from a kerosene rocket relatively low. A kerosene/LOX rocket produces an optimum vacuum I_{sp} of around 380 seconds, whereas for a LOX/LH₂ rocket the figure is 474 seconds (8). These I_{sp} figures and all quoted below are for the following conditions: P_c = 1000 psia, P_a = 0 psia, AR = 150, where P_c is chamber pressure, P_a is atmospheric pressure, and AR is the nozzle area ratio (A_{exit}/A_{throat}). Kerosene is best used when high thrust and ease of propellant handling, rather than performance, are the most crucial design factors.

Alternative Oxidizers. Hydrogen peroxide (H₂O₂) has been used as an oxidizer since the early days of rocketry. Decomposed hydrogen peroxide powered the turbopumps on the German V-2 (9). NASA has even used hydrogen peroxide in an auxiliary rocket engine for training. The NASA NF-104, used for training X-15 rocket plane pilots, was powered by a Rocketdyne AR2 engine that burned hydrogen peroxide and jet fuel. This engine used a catalyst to decompose the peroxide that was then used to power the turbine and was discharged

overboard. The engine was used for several flights until the X-15 program was discontinued.

Although liquid oxygen is by far the most common oxidizer for rocket engines, there are many other choices. Experimentation with cryogenic fluorine (F_2) as an oxidizer was carried out in the 1960s due to fluorine's extremely high performance characteristics. For expansion to a vacuum, an LF_2/LH_2 rocket can produce an I_{sp} nearly 20 seconds higher than a LOX/LH_2 rocket (8). Fluorine has a major drawback for use as a rocket fuel, however, and that is its toxicity. One of the products of combustion is hydrofluoric acid, a poisonous substance. Primarily because of this threat of pollution, fluorine rockets have never advanced past the developmental stage.

Storable Propellants. Uses for which propellants need to be stored on board the vehicle for long periods of time or when launch preparation times are necessarily short are good candidates for storable propellants. One early application of storable propellants arose on the Titan II vehicle. Engine development for this vehicle began in 1960. Originally developed as an ICBM, the Titan II could be launched in about 1 minute, compared to the 15 minutes required to launch its $LOX/RP-1$ powered Titan I predecessor. The Titan II was fueled by nitrogen tetroxide and Aerozine-50, propellants which were hypergolic as well as storable (10). The current configuration of the first stage engine, the LR87, is flown on the Titan IV vehicle. It develops 548,000 lb of vacuum thrust with a vacuum I_{sp} of 302 seconds (5).

Hypergols are common among storable propellants. In many cases, the engines must be ignited in space, must be operated numerous times, or used under other conditions that make highly reliable operation necessary. Elimination of the igniter removes one significant source of uncertainty from the hardware. The OMS engines and reaction control system for the Space Shuttle, as noted earlier, operate on nitrogen tetroxide and monomethyl hydrazine (MMH). This propellant combination has a higher bulk density than either LOX/LH_2 or $LOX/RP-1$, although its performance is modest. N_2O_4/MMH engines operating at a ratio of 2.37:1 generate an optimal vacuum I_{sp} of 341.5 seconds, lower than hydrogen- or kerosene-powered engines. Another combination of storable propellants uses nitrogen tetroxide as the oxidizer and a 50/50 mix of hydrazine and unsymmetrical dimethylhydrazine (UDMH). This combination, provides an I_{sp} less than 1 second greater than N_2O_4/MMH engines and has nearly the same bulk density (7). Most small-thrust, in-orbit applications, in which low performance is not a significant detriment and where it may be necessary to store propellants onboard the vehicle for an extended period of time, commonly use storable propellants.

Engine Performance

Optimization of the propulsion system implies minimizing the total amount of propellants that must be carried on the vehicle to satisfy the required thrust in all conditions across the vehicle trajectory. This typically is achieved by maximizing I_{sp} across the flight trajectory, which requires considering several design factors. The following discussion will consider the effects of chamber pressure

and the nozzle expansion ratio on the engine performance for applications ranging from sea-level launches to space operations in vacuum.

The net change in propellant velocity is a direct measure of the thrust produced per unit mass burned. Several factors influence the performance that can be achieved with an engine in a specific application. These factors include the selected propellant combination, the combustion chamber pressure, and the exit pressure to which the combustion gases are expanded. The propellant combination and the pressure at which they are burned determine the temperature and properties of the combustion products. The nozzle exit pressure is constrained either by the ambient atmospheric environment or by physical limitations of the spacecraft envelope. For launch vehicles, atmospheric pressure is the constraint, whereas at high altitude or in space, the vehicle envelope limits the allowable nozzle geometry.

The nozzle exit velocity V_e in Equation 2 is a function of nozzle exit pressure. For a constant propellant flowrate and a fixed exit area, as chamber pressure is increased, the throat area needed to pass the mass flow decreases and the nozzle area ratio increases, resulting in higher exit velocity. Therefore, the net effect of increasing the chamber pressure in an engine designed for a specified limiting envelope is an increase in specific impulse that yields a reduction in the fuel required to achieve a given level of thrust.

For application in booster systems, matching exit static pressure with local ambient pressure is typically the controlling restraint on nozzle design. It is most severe at launch and rapidly decreases with altitude. In practice, nozzles are designed for more than ideal expansion at launch (overexpansion) because this yields higher exit velocity and some gain in performance integrated over the flight trajectory. This is a useful design approach because it also reduces the amount of underexpansion that occurs at higher altitude. Care must be taken when designing overexpanded nozzles because too low an exit pressure results in flow separation within the nozzle. Separated flow is very detrimental because the flow is typically unstable and results in uncontrolled variations in the thrust direction. It can also have a significant negative effect on thrust.

Figure 1 shows the trade-off between area ratio and gross thrust (i.e., effects of external pressure not shown) for a LOX/hydrogen engine operating at sea level at a mixture ratio (O/F)=6:1, where combustion and nozzle performance are ideal and chamber pressure varies from 5–20 MPa. This figure, generated by NASA's Chemical Equilibrium Applications Program (11), shows the variation in impulse as a function of area ratio at the area ratio at which the nozzle exit pressure is 0.4 bar (note: 1 MPa = 145.03 psia; 1 MPa = 10 bar). This exit pressure is approximately the lowest pressure at which an axisymmetric bell nozzle can be operated at sea level without risk of internal flow separation. As shown in this figure, the final exit area ratio increases from 14 to 40 as the chamber pressure is increased from 5 to 20 MPa. The corresponding specific impulse increases from 397 to 433 seconds, indicating the benefit of initially increasing engine chamber pressure for sea-level launch vehicles.

Engine impulse efficiency is also an important parameter in defining performance. It is the product of the main chamber combustion efficiency and the primary nozzle expansion efficiency. Combustion efficiency measures how thoroughly the propellants mix and burn. Less than 100% efficiency means that some

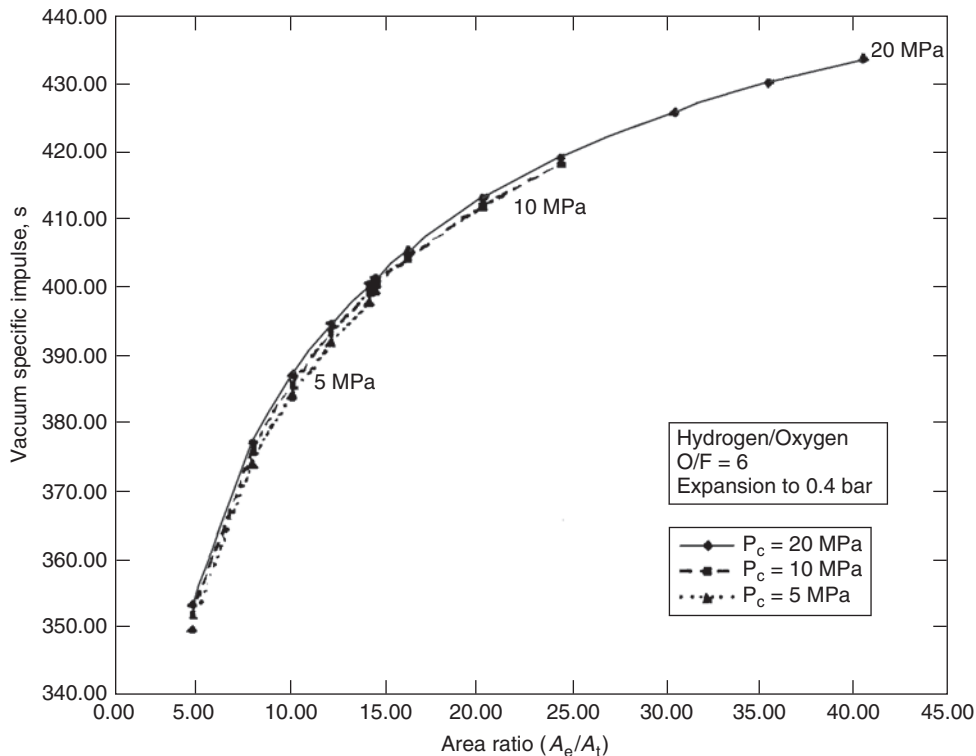


Figure 1. Variation of impulse with area ratio and chamber pressure.

of the fuel or oxidizer leaves the engine unburned, resulting in a loss of thrust. For oxygen and hydrogen propellants, the combustion efficiency for a well-designed system is generally 98 to 99%. For oxygen and kerosene propellants, the practical limit for combustion efficiency is generally 95 to 96%. Nozzle expansion efficiency reflects losses in thrust due to less than ideal expansion of gases in the nozzle. The primary reasons for the loss are gas friction on the nozzle surface and divergence of the flow at the nozzle exit; this means that not all of the flow is aligned parallel to the direction of thrust. Typical nozzle expansion efficiencies are generally from 98 to 99%, depending on several factors, including operating nozzle pressure ratio and the design area ratio.

Another way of looking at trade-off factors for engine performance is to examine the chamber pressure and specific impulse for expansion to a fixed exit pressure. First-order performance trends, as a function of chamber pressure and nozzle exit pressure, are shown in Fig. 2a, for hydrogen and Fig. 2b, for kerosene fuels. The theoretical specific impulse information in these figures was also generated by the NASA Chemical Equilibrium Program. This theoretical performance was adjusted according to the indicated constant values of combustion efficiency and nozzle expansion efficiency and further corrected to account for the external pressure on the engine. Performance is shown for three nozzle discharge pressures P_{exit} . The $P_{\text{exit}} = 1.0$ bar line represents the nozzle expansion ratios

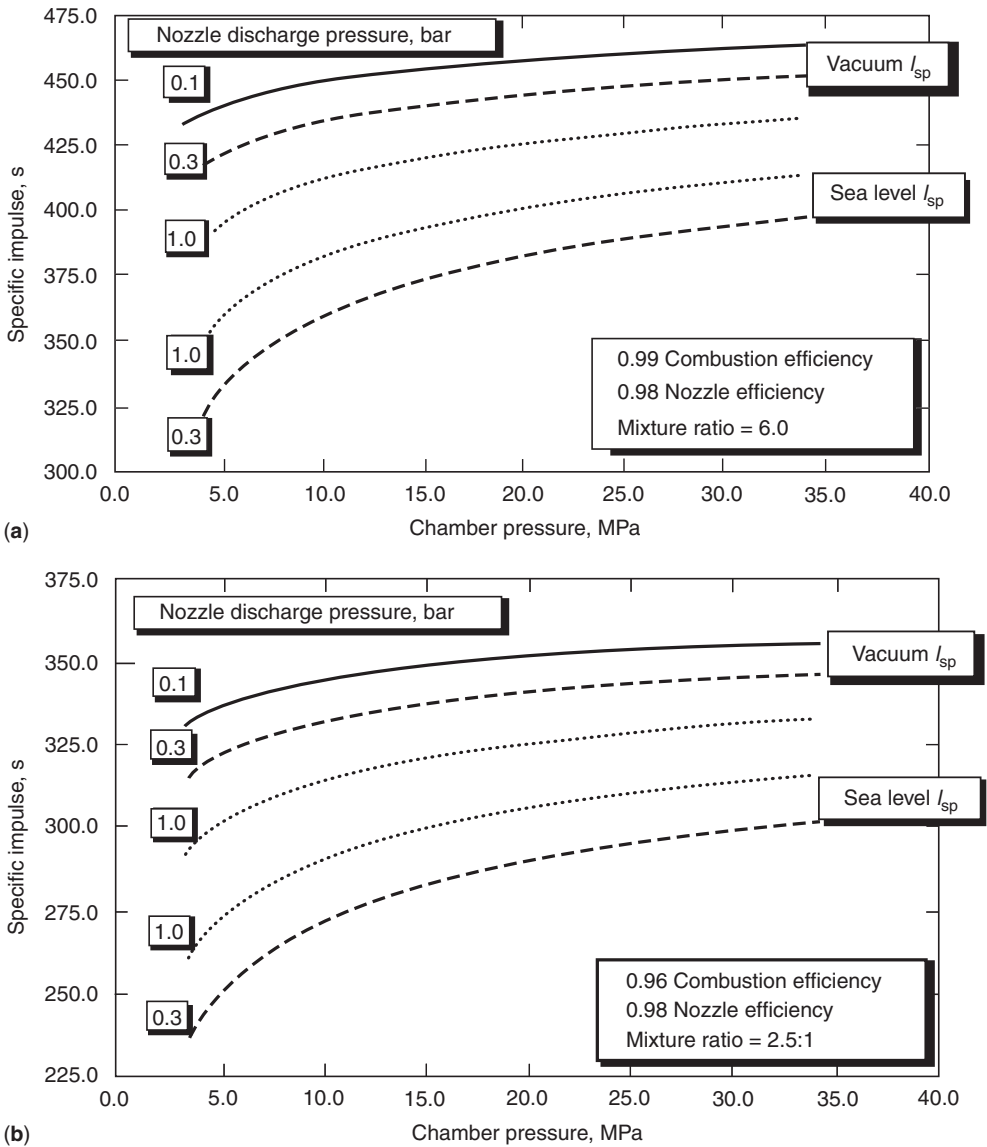


Figure 2. (a) Hydrogen/oxygen performance trends (courtesy R. Parsley, ONERA Proc., June 1995). (b) Kerosene/oxygen performance trends (courtesy R. Parsley, ONERA Proc., June 1995).

required to yield a sea-level atmospheric exhaust pressure. This is the exit condition generally required to maximize sea-level thrust. This is desirable and typical for booster applications. The $P_{exit}=0.3$ bar line represents the maximum sea-level expansion ratio that can be sustained without nozzle separation. This is desirable and typical for engines that must start at sea level but also operate at high altitudes. The $P_{exit}=0.1$ bar line is representative of upper stage expansion ratios that balance performance, weight, and engine size. The specific impulse

curves in Fig. 2a show that I_{sp} at sea level is reduced if the nozzle expands to a pressure below sea-level atmospheric pressure. This is due to the “negative” thrust effect of the pressure difference (internal minus external pressure) acting on the nozzle. This effect is predicted by Equation 5. In vacuum conditions, the effect of expanding to lower exit pressure is to produce higher performance due to the higher exit velocity produced. In this condition, the external pressure effect has disappeared, and the thrust due to propellant acceleration governs performance.

These performance estimates are presented only for initial screening. The secondary effects of changes in combustion efficiency and nozzle efficiency that depend on each individual design are important and should be investigated and optimized for each individual application.

It is apparent that increasing chamber pressure is beneficial for both launch vehicle applications and for space applications (limited by nozzle envelope constraints). So why not just increase combustion chamber pressures to reap the apparent benefits? Several factors weigh against this, including increased chamber cooling requirements, increased propellant pumping power, and increased chamber structural weight.

Engine Cycles

The objective in this section is to provide an overview of the thermodynamics that influences the configuration and optimization of liquid-fueled rocket cycles. The intent is to provide insight into the fundamental differences and inherent capabilities of different cycle approaches. Pressure-fed engines consist of a single cycle class. These are typically much simpler than pump-fed systems. Some of the issues specific to pressure-fed engines are covered in the section on Propellant Supply Systems.

Cycle Types and Configurations. Pump-fed liquid rocket cycles are defined by two configurational variables. The first cycle configurational variable is the energy source for the turbine drive. The turbine energy source can be from an auxiliary combustion device such as a preburner or a gas generator or from the main combustion chamber, either directly by extracting combusted propellants or indirectly by heat transfer through the chamber walls. The second cycle configurational variable is the turbine discharge location. Historically, there are two options for turbine discharge flow. If the turbine discharge is to a high-pressure region, specifically the main combustion chamber, the cycle is referred to as a “closed” cycle. If the turbine discharge is to a low-pressure region, generally overboard or into the nozzle skirt, the cycle is referred to as an “open” cycle.

Figure 3 is a summary of eight possible configurational options and includes the common names of each cycle. Also included are the options for turbine working gas supply, propellant limitations, and examples of operational engines of each cycle type. Five of the cycles have been developed into operational engines. This section examines the three most common cycle options. For simplicity, the supporting engine schematics do not include propellant boost pumps and are examined with separate turbopumps for the fuel and oxidizer. The schematics include the minimum valve complement required for engine start-up and control.

		Turbopump-Drive Energy Source		
		Main Combustor		Auxiliary Combustor
		Chamber Coolant Heat Limited to Hydrogen, Methane, Propane Fuels	Chamber Combustion Gas All Propellants are Compatible	Gas Generator or Preburner All Propellants are Compatible
Turbine Discharge Pressure Sink	High Pressure: Combustor Main	<u>Closed Expander</u> Fuel Cooling Fuel Cooling with Regeneration Fuel & Oxidizer Cooling (Full Flow) Operational Engine Examples: (RL10)	(No Flow Potential) (None Possible)	<u>Staged Combustion</u> Fuel-Rich Oxidizer-Rich Fuel & Oxidizer-Rich (full Flow) (SSME, LE-7, RD-170, RD-0120, NK-33)
	Intermediate Combustor Pressure: A/B	<u>Afterburning Open Expander</u> Fuel Cooling (None)	<u>Afterburning Tapoff</u> Fuel-Rich (None)	<u>Afterburning Gas Generator</u> Oxidizer-Rich Fuel Rich (None)
	Low Pressure: Overboard or in Nozzle Skirt	<u>Open Expander</u> Fuel Cooling (LE-5A)	<u>Tapoff Cycle</u> Fuel-Rich (J-2)	<u>Gas Generator Cycle</u> Oxidizer-Rich Fuel-Rich Fuel & Oxidizer-Rich (Full Flow) (F-1, J-2S, Vulcain, RS-27, LR87-AJ, YF-20)

Figure 3. Turbopump power options for pump-fed rocket engines (courtesy R. Parsley, ONERA Proc., June 1995).

The basic propellant supply approach for a pump-fed rocket is illustrated in Fig. 4. The propellants are increased in pressure using single- or multiple-stage pumps. A single- or multiple-stage turbine supplies the pump power. Because the propellants for an open cycle are pressurized only slightly above chamber pressure, pump work is minimized. A turbine pressure ratio of five or greater is possible because of the low-pressure exhaust. For a closed cycle, the turbine drive flow is discharged into the main chamber, which is at a relatively high pressure. This generally limits the turbine pressure ratio to two or less to avoid excessive pump discharge pressures. For either the open or closed cycle approach, it is necessary to introduce energy into the turbine working fluid before expansion through the turbine. Depending on the option selected to provide this turbine energy, the cycle definition is different. The three common thermodynamic cycles for liquid rocket engines are expander, gas generator, and staged combustion.

The expander cycle, Fig. 5a, is a cycle in which hydrogen, or some other fuel, is used to cool the thrust chamber and nozzle regeneratively. Thermal energy absorbed during cooling of the chamber and nozzle heats the hydrogen fuel. The heated hydrogen, now in a gaseous state, passes through turbines and powers the pumps. In this engine cycle, the turbine gases are routed to the injector and main chamber, where they are combusted and expanded through the nozzle. The thrust chamber and nozzle heat transfer limits the energy available for the expander cycle. This limits chamber pressure potential to about 10 MPa (1500 psia).

Its simplicity is the major benefit of this cycle. There is no subsystem needed to provide the energy to drive the turbines. By the same token, its main drawback is that the energy of the turbine working fluid is limited and leads to relatively low chamber pressure.

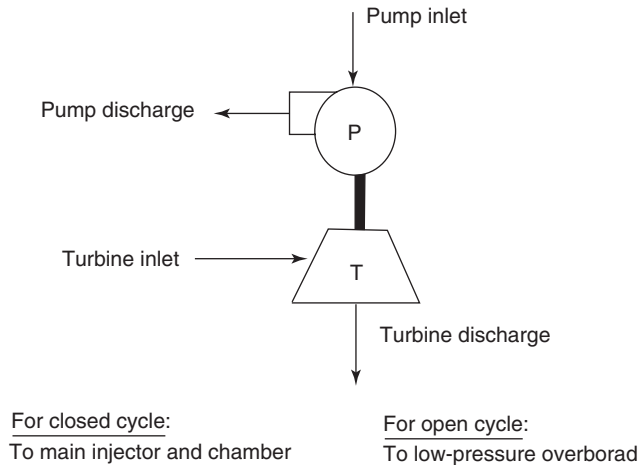


Figure 4. Pump-fed propellant supply schematic.

The gas generator cycle, Fig. 5b, is an open cycle configured so that a minimum fraction of the propellants is delivered to the gas generator combustion device before being directed to the high expansion ratio turbine. The pressure of the turbine discharge is less than the main combustion chamber pressure, and therefore this flow must bypass the main combustion chamber. The gas generator cycle dumps the gases used to power the turbopumps overboard or into the divergent section of the nozzle. The chemical energy released during combustion in the gas generator is restricted by the temperature limit of the turbine. Chamber pressure for a gas generator cycle is selected to optimize total engine

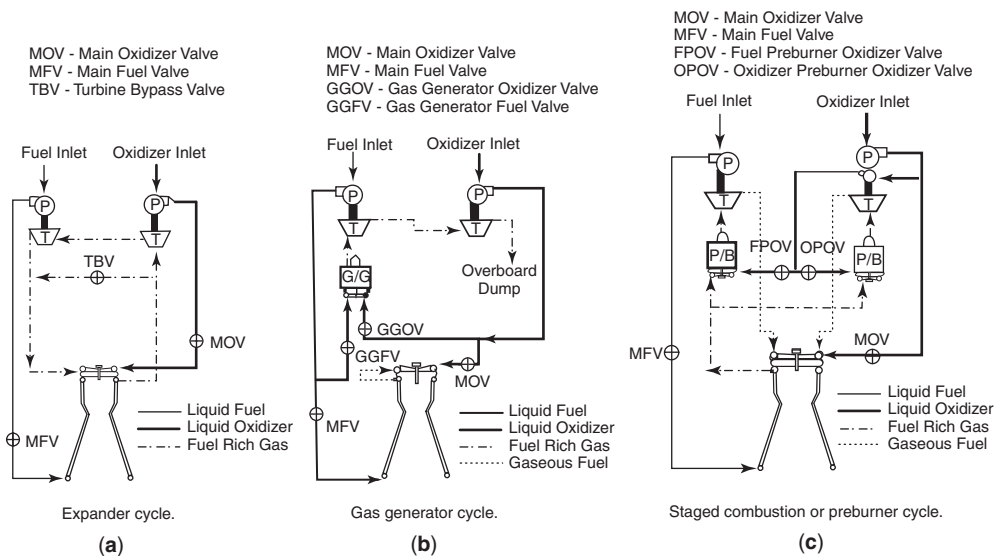


Figure 5. Schematics of liquid rocket engine cycles: (a) Expander cycle. (b) Gas generator cycle. (c) Staged combustion or preburner cycle.

performance, which includes both the higher performance main engine flow and the lower performance turbine discharge flow. This performance optimum generally occurs at 10 to 17 MPa chamber pressure, depending on propellant selection, and the overboard flow is generally less than 4% of the total engine flow.

The propellants for the gas generator are typically the same as those burned in the main chamber, though decomposition of hydrogen peroxide has sometimes been used, most notably in the German V-2 (9). The Rocketdyne RS-68 gas generator engine recoups some of its performance loss by using liquid hydrogen as a fuel, but its 410-second average vacuum specific impulse is still much lower than that of a comparable staged combustion engine.

The use of gas generators allows increasing the chamber pressure above that which is possible in the expander cycle. The cost of this is added complexity and some loss of thrust from dumping turbine discharge gases.

The staged combustion cycle, Fig. 5c, is a closed cycle configuration such that portions of the propellants are burned fuel-rich in preburner combustion devices upstream of the turbines. This heated mixture of fuel and combustion products is expanded through the turbine and fed into the main combustion chamber. In the SSME, the primary operating example of this cycle, approximately 80% of the fuel flows through the fuel turbine. The system must be balanced between the desire for high chamber pressure and the need to limit turbine inlet temperature to an acceptable value dictated by hardware requirements. Turbine inlet temperature is controlled by the amount of oxidizer that is fed into the preburner. For the SSME example, the mixture ratio in the preburners is of the order of unity. The performance of the staged combustion cycle begins to become hardware-limited between 20 and 24 MPa (3000 and 3500 psia) chamber pressure.

Newer designs would probably achieve pressures near 27 MPa. Use of full flow, ultrahigh performance designs, where a low mixture ratio, fuel-rich flow drives the hydrogen pump, and a high mixture ratio, oxygen-rich flow drives the oxidizer pump, can achieve thrust chamber combustion pressures in the 27 to 38 MPa (4000 to 5500 psia) range. Use of cooled preburner and turbine hardware (including turbine blades) can extend the combustion pressure to 55 MPa.

Table 2 illustrates some of the basic cycle parameters for a variety of liquid rocket engines in current use, including thrust chamber operating conditions, type of cycle, and propellant pump conditions.

Propellant Supply Systems

The propellant supply system consists of the various components that store the propellants in the vehicle and deliver them in a controlled manner to the engines. These components include the propellant tanks, valves, pumps, plumbing, and pressurization subsystem. Depending on the type of propellant pressurization (i.e., pumped or pressure-fed), the propellants are supplied as liquids from the vehicle tankage to the inlets of the main pumps or to valves that control the flow into the engine.

Table 2. General Cycle Characteristics of Some Current Rocket Engines

Engine	Vehicle	Propellants	Cycle	P_c , psia	T_c , °R	Pump, hp	Pump speed, rpm
SSME (Blk II)	Shuttle	LOX/H ₂	Staged combustion	3000	6500	69,000 fuel, 25,150 LOX	35,000 fuel, 24,000 LOX
RL10A-4 (Pratt Whitney)	Atlas-Centaur	LOX/H ₂	Expander	565	6100	740 fuel, 160 LOX	35,500 fuel, 14,200 LOX
Vulcain (SEP)	Ariane 5	LOX/H ₂	Gas generator	1600	6300	16,100 fuel, 5,100 LOX	33,500 fuel, 13,800 LOX
RD170 (NPO Energomash)	Zenit	LOX/RP-1	Staged combustion	3500	6900	55,000 fuel (Main + kick), 123,00 LOX	17,000 both (single shaft)
RS-27A (Rocketdyne)	Delta II, Delta III	LOX/RP-1	Gas generator	700	6400	2000 fuel, 1100 LOX	6800 fuel, 6800 LOX
LR-87-AJ-11 (Aerojet)	Titan IV	N ₂ O ₄ /Aerozine 50	Gas generator	860	5900	3800 fuel, 4000 LOX	9500 fuel, 8700 LOX

The parameter conventionally used to quantify the inlet condition is the net positive suction head (NPSH), the total head difference (i.e., pressure plus velocity head) between the propellants at the inlet and their corresponding vapor pressure. It is desirable that the NPSH at the pump inlet be sufficient to avoid cavitation in the pump. One way to achieve this is by maintaining high propellant tank pressure. This approach may result in excessively heavy tanks, particularly for propellants with high vapor pressures. In addition, the empty tank volume, called “ullage,” must be at the same pressure as the propellants in the tank. As liquid is withdrawn from the tanks, pressurized gas must be injected to replace the liquid volume. This ullage gas can be either heated and vaporized propellant or an inert pressurized gas. At higher tank pressures, this represents a significant mass of pressurizing gas, particularly as the propellant level in the tank approaches “empty.”

One design option is to supply the propellants at pressures that are a compromise between minimizing the tank ullage pressures and maximizing the net positive suction head. The vehicle prefers low ullage pressure so that the tank walls are thinner and therefore lighter in weight. This creates a lighter or smaller vehicle for the same propellant load and translates into larger payloads delivered to orbit. Alternatively, for a given engine operating chamber pressure, pump-fed engines need high enough inlet pressures to prevent cavitation in the pump. Selecting a propellant with a vapor pressure too low or too high can severely affect the overall tank and propellant feed supply design. It could require higher ullage pressures that lead to heavier tanks on the vehicle.

A second option is to design the tanks for low ullage pressures to minimize tank weight and to satisfy the required NPSH level by adding propellant boost pumps to increase the net inlet pressure to the main pumps. This adds complication to the system but can provide significant overall benefit to the vehicle. Boost pumps are typically small axial flow pumps designed to operate at low pressure, referred to as “inducers”; they are placed either at the tank or at the inlet of the main pump. The inducer has the effect of increasing the net positive suction head at the main impeller inlet by boosting the total fluid pressure.

Pressure-fed rocket engines can be designed to supply a single propellant (i.e., hydrazine thrusters) or as a bipropellant system, like the space shuttle OMS. Tank pressurization is achieved via complementary pressurization tanks or bottles of an inert gas such as helium or nitrogen. Regulator valves are used to control the gas pressure relative to the fluid pressure in the main tanks. The system must be designed to compensate for the pressure losses between the pressurization tanks, valves, main tanks, propellant lines, and the main chamber injector. Pressure-fed propellant supply systems are normally limited to engine operating pressures of less than 3 MPa.

The propellants are supplied to the engine inlet valves, control valves, or pump inlets via fixed or flexible propellant feed lines. In some engine designs where vernier or auxiliary engines provide steering the engine does not move. In that case, the engine propellant supply lines are usually hard-mounted to the inlet valves or pumps. When the engine is vectored or gimballed to provide flight control, the propellant supply will use a flexible duct line that permits “jointed” movement in two planes. In some cases (e.g., Russian RD-170 LOX/kerosene

engine), the propellant supply lines are fixed, and the thrust chamber gimbals to provide vectored flight control.

Liquid Propellants Turbopumps

The function of the rocket engine turbopump is to receive the liquid propellants from the vehicle tanks at low pressure and supply them to the combustion chamber at the required flow rate and injection pressure. The energy to power the turbine is provided by the expansion of high-pressure gases that are often mixtures of the propellants being pumped. This section relies heavily on Reference 12.

The liquid rocket engine turbopump is a unique piece of rotating machinery. The turbopump typically pumps cryogenic liquids and is driven by high-temperature gases, creating large temperature differentials between the pump and turbine. The pump must avoid cavitation while pumping relatively high-density fluids at low inlet pressures and deliver them to the thrust chamber at very high pressures across a relatively wide throttling range. The turbine is often driven by fuel-rich combustion products that have very high available energy and heat capacity levels. The turbopump is optimized for performance and weight within the minimum possible envelope size to facilitate engine packaging. The bearings normally operate in the environment of the propellants being pumped, which have minimal lubrication characteristics. The static and dynamic seals must preclude mixing propellants within the turbopump, which would result in burning and catastrophic failure.

Engine Requirements. The type of engine cycle has the most significant influence on the turbopump requirements and configuration. Other major engine factors that significantly influence the turbopump configuration are the types of propellants, the propellant inlet conditions, and the engine throttling requirements. Variations in density produce significantly different pump head rise requirements and large differences in volumetric flow, that is, low-density propellants require a much higher head rise to develop the same discharge pressure (head rise = pressure rise/density, $\Delta H = \Delta P/\rho$). The variations in the combined propellant available energy also have a significant influence on turbine design.

The pump suction performance requirement is its ability to operate at the available NPSH without cavitation sufficient to affect its ability to develop the required discharge pressure and flow rate.

The engine throttling requirements define the range of flow and discharge pressure that the turbopump must deliver in stable operation. The engine start and shutdown characteristics must also be considered to prevent unstable turbopump operation caused by cavitation or stall.

When the engine requirements are established, the turbopump configuration is selected based on optimizing the pumps for each propellant, the turbine for the drive gas available energy, and the mechanical design arrangement for life, weight, and producibility considerations. Maximum pump speed is generally limited by the suction performance requirements to avoid cavitation. The optimum turbine speed for maximum efficiency and minimum weight is generally

higher than the high-density fluid pump speed. Maximum turbine efficiency requires a certain pitch-line velocity, which is the product of the shaft speed and the turbine diameter. The minimum weight turbine has the highest speed and smallest diameter within the structural and mechanical arrangement limitations.

Earlier engines with small power requirements sometimes employed a gearbox to match the speeds of the pumps and turbine better, but at the very large power levels of launch systems, pumps are typically direct driven by the turbine. Therefore the turbine must satisfy the power requirements of the pump at the same shaft speed.

Pumps. Inlet conditions (NPSH), discharge pressure, flow rate, and operating range must all be satisfied by the pump configuration. A parametric analysis is performed to select the best speed, diameter, and number of stages compatible with the turbine and mechanical design considerations (Fig. 6).

The pump inlet diameter selected is generally based on the available NPSH. Test experience has been accumulated on inducers to correlate their suction performance as functions of the NPSH, the fluid inlet meridional velocity C_m , and the inducer flow coefficient ϕ . The inducer diameter (inlet area) is selected to limit the fluid meridional velocity so that the available $NPSH/2g_cC_m^2$ is equal to

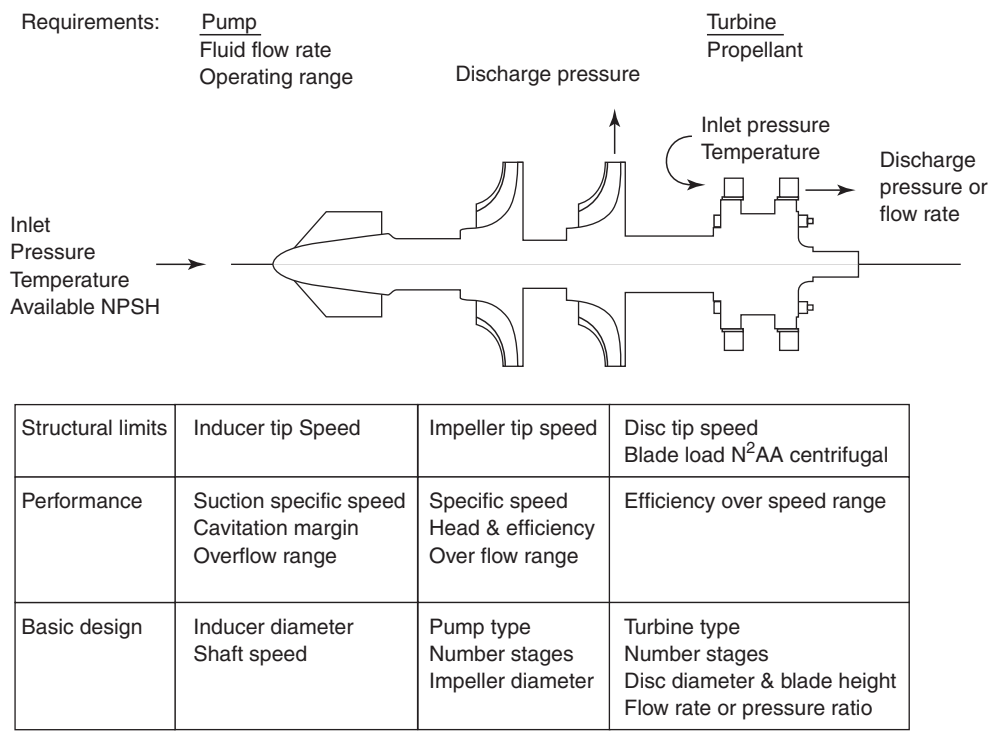


Figure 6. Performance and structural limits establish basic design.

or greater than three velocity heads for water, two for LOX and one for LH_2 . The variation in the empirical limit accounts for the difference in the thermodynamic suppression head among water, LOX and LH_2 . The limit is also a function of the inducer flow coefficient, which is defined as the meridional velocity divided by the inducer tip speed, $\phi = C_m/U_t$.

When the inlet diameter is selected, the shaft speed is selected to limit the inducer tip speed to approximately 550 ft/s in LOX and 1100 ft/s in LH_2 . The tip speed limit is for controlling the tip vortex cavitation energy. The blade thickness must also increase with increased tip speeds to react to the centrifugal and pressure loading. This results in reducing the flow passage area and, therefore, lowers the suction performance. The pump suction specific speed is expressed as $S_s = N\sqrt{Q}/(\text{NPSH})^{3/4}$, which is a measure of the pump's ability to operate at low inlet NPSH without cavitation sufficient to cause head loss. A 50% NPSH margin is generally selected during the design process for long-life rocket engine applications. Cavitation, in addition to decreasing the pump discharge pressure and efficiency resulting from the formation of vapor bubbles, can cause significant structural damage when the vapor bubbles collapse (implode), particularly in high-density fluids. Inducer technology development has been a key state-of-the-art advancement for increasing pump speed, decreasing turbopump weight, and increasing safe operating life (Fig. 7).

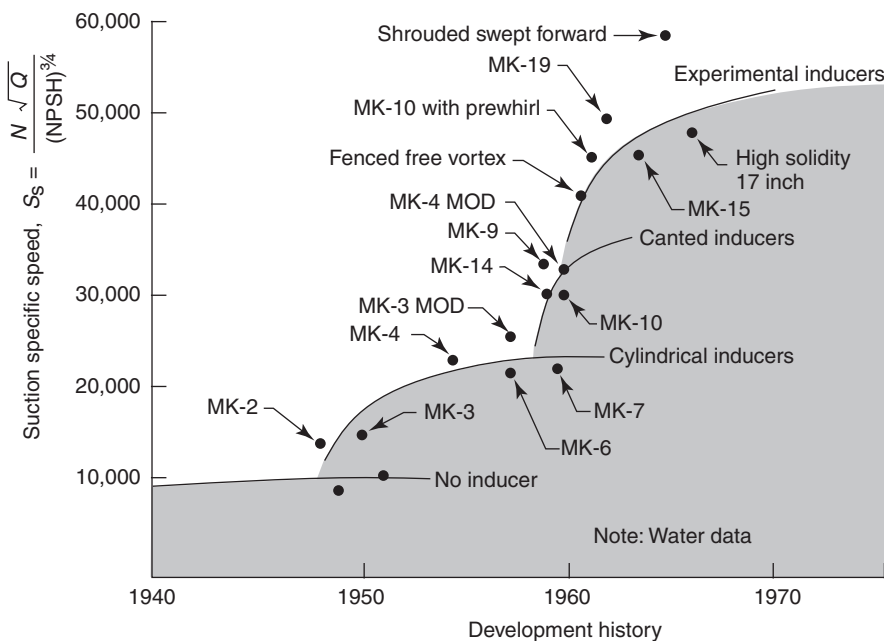


Figure 7. Rocketdyne suction specific speed history.

Required pump head, which is a function of the required discharge pressure, the available inlet pressure, and the propellant density [$\Delta H = (P_d - P_{in})/\rho$], is the major factor in selecting the pump configuration. The head coefficient is a function of the pump type and establishes the required pumping element diameter and number of stages for a given shaft speed. The main pumping element may be a centrifugal, mixed, or axial flow type (Fig. 8).

For the combination of very high flow rate and large head rise usually encountered in moderate and large engine applications, centrifugal pumps are typically chosen as the most appropriate type of pump. In this pump type, pressure rise is achieved by a combination of acceleration and centrifugal force, as the fluid flows along a curved flow path. The pressure rise achieved depends on both the characteristics of the pump (i.e., impeller design and rotational speed) and on the density of the fluid. In general, pressure rise in a single centrifugal stage is proportional to the product of the fluid density and the square of the tangential velocity at the outer rim of the impeller (tip speed). Thus, the higher the density, the lower the required tip speed to achieve a given pressure rise. In general, the tip speed and impeller diameter are well within aluminum and nickel-base alloy structural limits. The head requirements for low-density fluids, such as LH_2 , are very high and typically require several stages to develop. An axial flow main pumping element was selected for the J-2 LH_2 pump because of its intermediate specific speed and the narrow throttling range requirements. The 200,000-ft

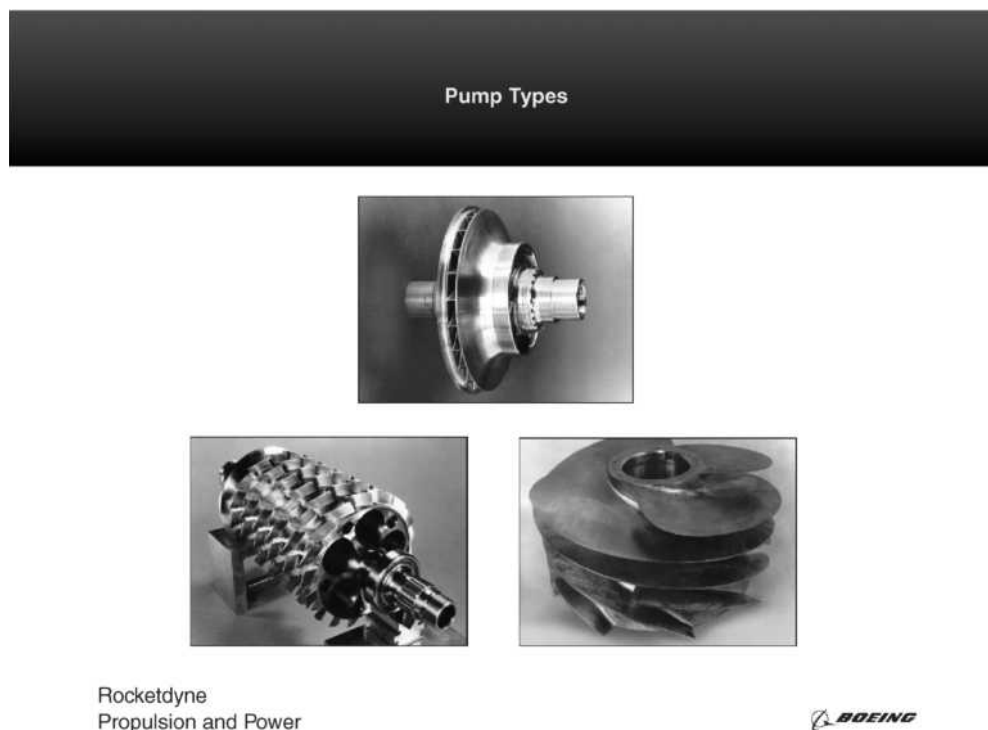


Figure 8. Pump types (photos). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

head requirement for the SSME high-pressure fuel turbopump dictated a three-stage centrifugal pump with impellers operating at 2,000 ft/s tip speed. Titanium, which has a higher strength-to-weight ratio than high-strength nickel-base alloys, was required by the high tip speed.

The pumping power requirement depends on the flow rate, pressure ratio, and thermal efficiency and can be calculated from

$$HP = w_p(H_{o,actual} - H_i) = w_p(H_{o,ideal} - H_i)/\eta, \quad (6)$$

where w_p is the propellant flow rate, H_i is the inlet enthalpy, $H_{o,actual}$ is the actual exit enthalpy, $H_{o,ideal}$ is the ideal isentropic exit enthalpy, and η is the pump efficiency. The efficiency is determined by the internal aerodynamics of the pump and is typically characterized for a defined pump geometry as a function of speed and pressure ratio or speed and flow rate. Pumps can be designed to operate at high efficiency at their design conditions, but efficiencies tend to be lower if the pump is operated at conditions far from design. It is desirable to design the pump to match best the cycle conditions over the engine operating envelope, maximizing the efficiency to minimize power requirements.

The Space Shuttle Main Engine (SSME) represents the extremes of current turbopump requirements. The design requirements of the liquid hydrogen fuel pump at maximum power are 162 lb/s flow rate at a discharge pressure of 6400 psia. The liquid oxygen pump delivers 1161 lb/s at 7300 psia at maximum power. It is noteworthy, on examining the operating characteristics shown in Table 2, that hydrogen fuel represents about 12% of the total propellant flow but requires about 74% of the total pumping power. This is the result of the relatively low density of liquid hydrogen. The table also shows that the speed of the hydrogen pump is 50% greater than that of the oxygen pump, again because of the very large head rise necessary to produce the desired pressure rise in a low-density fluid.

An ideal fluid with regard to pumping requirements would be inert with respect to the pump materials, would have a relatively high density, and would be a good lubricant and coolant for use in the pump bearings. To illustrate the contrast in propellant characteristics, consider some of the propellants in current use (Table 3).

Optimizing the pump efficiency, which is a measure of the work out/work in, can also influence the shaft speed and specific speed selected. Figure 9 shows the relationship between efficiency and specific speed. Small flow rate pumps are generally less efficient than large flow rate pumps because the clearance and surface-finish-related losses cannot be scaled with size.

Turbines. The turbine must supply the required power to drive the pump, using the drive gas provided by the selected engine cycle. It is desirable that the working fluid be at the maximum possible temperature to provide the maximum work potential. However, the turbine inlet temperature is limited by the material and structural capabilities of the turbine because it is very desirable to avoid cooling the turbine. Axial flow turbines tolerate high gas temperatures better than radial inflow turbines and generally have much lower thermal stresses. For this reason, turbines for rocket engines are generally the axial flow type and may use either impulse or reaction type aerodynamics. Small engines will usually use

Table 3. Propellant Characteristics of Importance for Pumps

RP-1 (Kerosene)	Hydrogen	Oxygen
Inert with regard to most pump material	Causes embrittlement in nickel alloys because of chemical reaction	Highly reactive with many materials
Ambient temperature storable	Deep cryogen, liquid at 20 K Materials must be cryogenic-capable (not brittle at hydrogen temperature)	Cryogenic, liquid at 90 K Materials must be cryogenic-capable and resistant to oxidation
Relatively good lubricant	Virtually no lubrication capability, although very good coolant	Poor lubricant qualities high density ~ 71 lb/ft ³
Density ~ 50 lb/ft ³ Nearly ideal for ease of pumping	Density is very low ~ 4.4 lb/ft ³ Difficult propellant for pump design	Easier pumping than hydrogen, but requires careful material choices

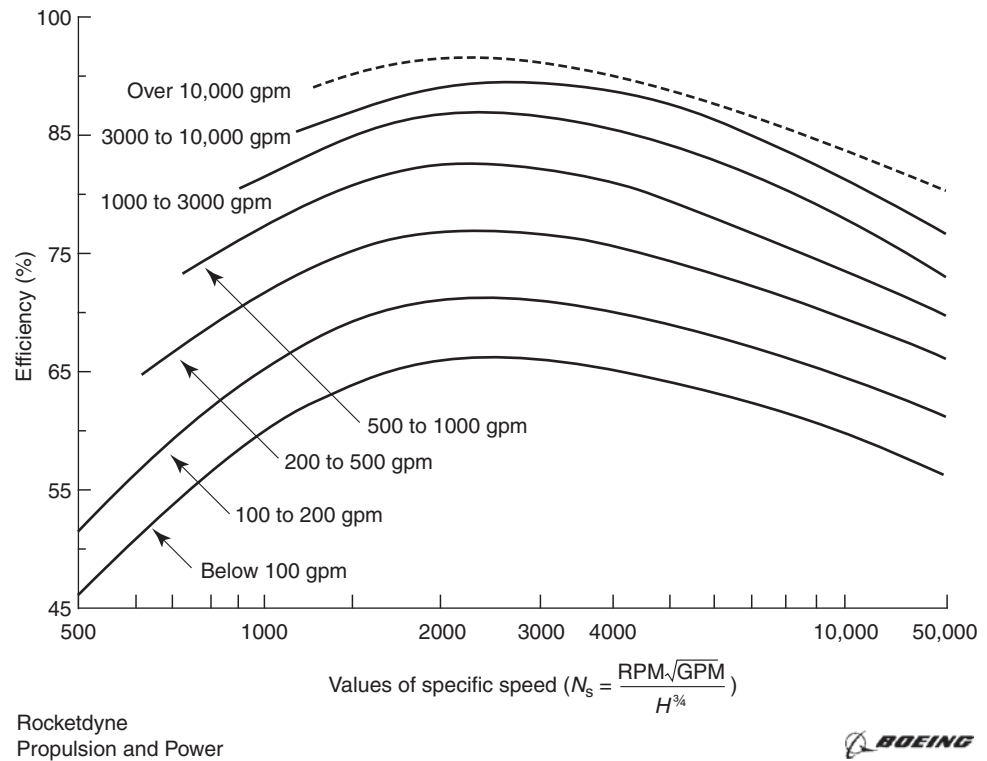
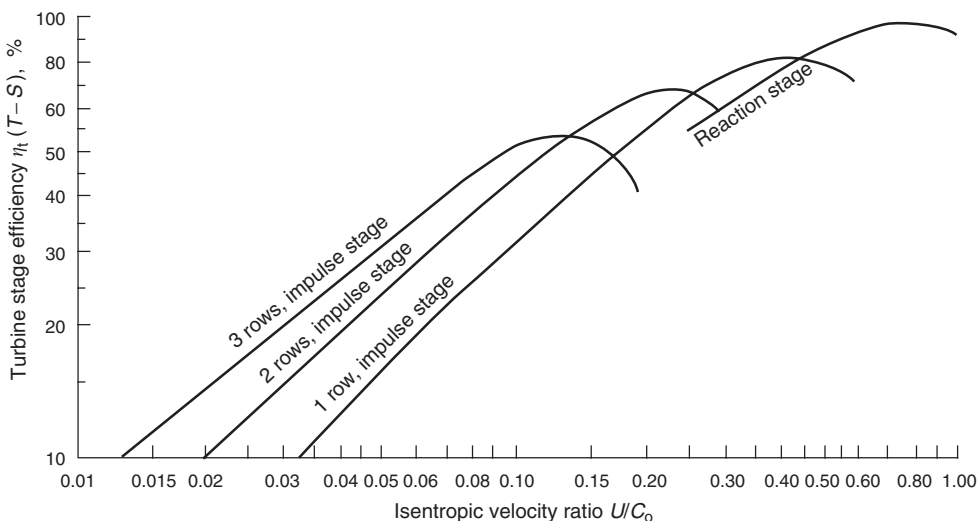


Figure 9. Variation of pump efficiency with specific speed.

impulse turbines because of their ability to use partial admission nozzles and avoid significant leakage losses. Large thrust engines may use reaction turbines, if the working flows are large enough to avoid requirements for small blading. Small reaction turbines are difficult to use because they require extremely tight clearances and seals to avoid significant flow leakage and have large attendant performance losses.

The operating characteristics of turbines are defined by the power produced as a function of the flow rate and the pressure ratio across the turbine. This is determined by the product of the flow rate and the enthalpy change as the working fluid expands in the turbine. This expansion process takes place as flow is accelerated across the nozzles or vanes and converts pressure to kinetic energy, and momentum is removed as forces acting on the rotating blades. Another way of looking at turbine performance is that it depends on three variables: the available energy content of the gas, the blade tangential velocity U , and the number of turbine stages. The available energy for the turbine pressure ratio can be expressed as an ideal velocity C . The turbine velocity ratio U/C is used to characterize these two variables empirically versus the turbine efficiency.

The ideal velocity can be distributed among the turbine stages by either a pressure-compounded or velocity-compounded design. The major difference between these two turbine designs is where the expansion occurs in the stationary blade rows. For the velocity-compounded turbine, all of the expansion occurs in the first stationary blade row, and for a pressure-compounded turbine, the expansion is distributed between the stationary blade rows. For high U/C designs, the turbine efficiency can be further improved by having some of the expansion (reaction) take place in the rotor blades, as illustrated in Fig. 10. The design selection is made to maximize the turbine efficiency and minimize the weight,



Rocketdyne
Propulsion and Power



Figure 10. Velocity ratio vs. efficiency for impulse and reaction staging.

compatible with the selected shaft speed. In general, when a direct-drive turbopump configuration is selected, the shaft speed is less than optimum for the turbine, and additional stages must be added to use the available energy.

The blade tip diameter is selected to optimize the U/C for efficiency within the blade height-to-diameter performance limits and within the tip speed structural limits. If the blade height-to-diameter ratio becomes too small, the tip clearance and secondary flow losses become large and decrease the turbine efficiency. The tip speed structural limit is based on the centrifugal pull that can be carried at the base of the blade airfoil for the selected material. Partial admission turbines are selected when the shaft speed is too slow and the blade height-to-diameter ratio becomes too small to obtain the desired U/C . The blade diameter is enlarged to increase U , and the arc of admission is decreased to maintain the blade height at an acceptable height-to-diameter ratio.

Mechanical Design. The mechanical design is a compromise that involves many contributing factors. Major factors that influence the mechanical design are power transmission, rotor dynamics, axial thrust balance, selection of bearings and dynamic seals, and thermal considerations.

The shaft diameters, splines, and couplings must all be large enough to transfer the torque, which is a function of the speed and horsepower. This establishes the shaft diameters and the minimum allowable diameter of the bearing inner race, depending on the axial stations selected along the shaft. Rotor critical speeds are a function of the rotor mass distribution and stiffness, bearing locations and spring rate, and the housing stiffness. For rolling element bearings, the product of the inner diameter and the shaft speed (DN) is used as a measure of the bearing internal loading. Empirical life limits have been established for DN as a function of the propellant or lubricant used to cool the bearing. Based on these interacting limitations, the bearing locations are selected to keep the operating speed range clear of critical speeds and minimize the bearing bore diameter to maximize bearing life.

From the standpoint of critical speed, inboard bearings decrease the bearing span and increase the first critical speed, so long as the overhang does not exceed approximately one-half the span length. From a bearing standpoint, the most desirable location is outboard, so that the bearing size and geometry can be optimized independent of the required shaft diameter. Bearings outboard of the turbine generally require additional cooling, dynamic seals, and support structure, which add complexity to the turbopump design.

Rotor stability is also a major factor in selecting the bearing locations and types of dynamic seals. Additional support stiffness and damping can be provided by the dynamic seals to raise the critical speeds and stabilize the rotor to prevent subsynchronous whirl. The rotor axial thrust is the other major factor that influences bearing design. The labyrinth seal diameters in the pumps and turbine are selected to minimize the net rotor thrust to which the bearings must react.

The rocket engine turbopump, in addition to being a high energy/weight ratio machine, must be designed to operate with the pump at cryogenic conditions and the turbine at high temperature. This requires design concepts that provide thermal growth flexibility while reacting to large torques and separating loads.

Materials. Aluminum alloys, stainless steels, high-strength steels, nickel-base alloys, cobalt-base alloys, and titanium alloys are all used in the design of

rocket engine turbopumps. Complex pressure vessels for applications up to approximately 2000 psi are typically cast of aluminum to use its high strength-to-weight ratio and to avoid welded joints. Nickel-base superalloys, such as Inco 718, are used to cast pressure vessels when higher strength is required. The high strength/weight ratio of titanium is used to obtain the high tip speeds required for LH_2 impellers and inducers.

The embrittling effects of gaseous hydrogen limit the materials suitable for turbine components. High-strength superalloys typically must be protected from the environment by copper or gold plating. Turbine blades are directionally solidified and thermally coated to survive heat fluxes 10 times the typical turbojet and blade loads up to 600 hp per blade.

Silver and Kel-F are used in LOX pumps where contact with the inducer or impeller could result in ignition caused by local heat generation. These materials are also used for potential contact with titanium impellers to preclude formation of titanium hydrides caused by heat generation.

Probably the most significant technological advancement to impact future turbopump designs is the development of fluid film bearings. Hydrostatic bearings remove the DN constraints of rolling element bearings, increase the bearing direct stiffness by a factor of 5, and increase damping by a factor of 100. Rotor stiffness can be increased, and the bearings can be positioned to optimize the machinery arrangement for rotor dynamics and performance considerations. A significant technological need for LH_2 turbopumps is the development of high strength-to-weight materials. Specific strengths higher than titanium will increase the 2000-ft/s tip speed limit, increase the head generated per stage, and decrease the number of stages required for the desired discharge pressure. Further information on this subject can be found in Reference 13.

Thrust Chamber Design

A thrust chamber is that part of the rocket engine that produces the thrust; it is defined as the section of the flow path enclosure extending from the injector face to some location downstream of the nozzle throat, at which the thermal and structural loads are reduced sufficiently to allow significant reduction in the weight of the flow path structure. The nozzle structure downstream of this location is usually referred to as the nozzle extension. The thrust force is transmitted to the vehicle thrust frame through a thrust structure mounted on top of the injector that allows gimbaling the whole engine. Figure 11 shows a typical thrust chamber configuration found in all flight engine designs today, using an axisymmetric chamber and a bell-shaped nozzle extension. Other configurations such as circular or linear aerospike designs or expansion-deflection nozzle designs have been successfully used in experimental engines but have not matured enough to demonstrate performance benefits in flight hardware compared to the more traditional design.

In the thrust chamber, the propellants are manifolded and injected into the combustor by the injector and are atomized, mixed and burned inside of the combustor. The resulting hot gases are expanded inside the nozzle extension, accelerating the flow to high exhaust velocities and producing the required



Figure 11. Photo of regeneratively cooled thrust chamber (Vulcain). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

thrust. In a typical design, about half the thrust is generated by the flow downstream of the thrust chamber throat. The thrust chamber is functionally characterized by components that provide stable and highly efficient propellant combustion and efficient combustor and nozzle cooling. Design activities at the thrust chamber level start out with design trade-offs to define optimum thrust chamber performance and resulting component requirements. Typical system interface issues that must be covered include the injector–combustor interface (“chamber wall compatibility”), trading combustion efficiency and stability versus chamber wall heat flux loading, and the location of the combustion chamber–nozzle extension interface. This usually involves trading off the system weight against performance and manufacturing requirements. For nonhypergolic propellant combinations, such as LOX/RP-1 and LOX/LH₂, an igniter is needed to initiate combustion. In most cases, an electrical ignition system is employed. The thrust chamber design must take into account all loads encountered during steady and engine start and stop operation, such as pressure loads, thermal loads, nozzle side loads, as well as taking into account requirements for external loads and geometric interface locations.

Figure 11 shows the cross-section in a photograph of a 1600-psia injector/combustor assembly, which will be used to outline the major design issues and features of a regeneratively cooled high-pressure thrust chamber. The design characteristics shown are similar for all hydrogen/oxygen systems used worldwide, such as the SSME (U.S.), the RD-120 (Russia), the Vulcain (Europe), and the LE-7 (Japan). Most regeneratively cooled chambers are fuel-cooled in a counterflow cooling arrangement; the regenerative section of the nozzle extension is also fuel-cooled. For overall engine performance, it is important that the coolant pressure drop is minimized.

The main design goals and requirements for the injector are high combustion efficiency and stability, providing additional chamber wall cooling, if needed, and sufficient structural strength and life. This leads to the appropriate definition of the geometry (e.g., the manifold shapes), the injector element design, the element pattern, and the materials and manufacturing processes. High combustion efficiency requires a uniform mixture distribution across the injector face and fine propellant atomization. A suitable manifold achieves the uniform mixture distribution and injection pattern design. The selection and optimization of the most suitable element type provides fine propellant atomization. Coaxial injection elements are used for liquid/gas systems like the one shown in Fig. 12 (LOX/GH₂) and for gas/gas systems. Liquid/liquid systems such as LOX/RP-1 or MMH/N₂O₄ usually feature impingement type injection elements, even

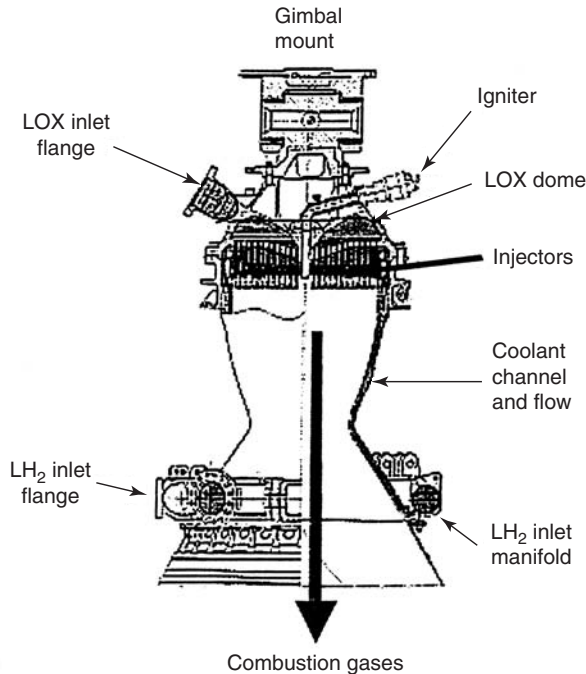


Figure 12. Schematic of Vulcain regeneratively cooled thrust chamber (courtesy of DASA).

though coaxial element designs are also successfully used in MMH/N₂O₄ systems. Unfortunately, many of the injection parameters that provide high performance, such as fine atomization, tend to reduce the combustion stability margin, and stable operation must be achieved by damping the acoustic processes (i.e., resonator cavities in the combustor wall) or by detuning them (i.e., baffles in the injector). Storable propellant combinations like MMH/N₂O₄ are most susceptible to combustion instability problems, LOX/RP-1 less so. The hydrogen/oxygen propellant combination is the least susceptible to high-frequency instability. To avoid coupling the combustion process with a feed system hydraulic mode ("chugging"), a sufficiently high pressure drop must be designed into the injector element, typically about 20% of the chamber pressure.

A typical injector like the one shown in Fig. 11 has about 500 injection elements. The faceplate needs to be actively cooled; this is achieved either by transpiration or regenerative cooling. Most parts of the injector are made from high-strength materials such as Inconel 718 or titanium, and for lower pressure applications, aluminum. The main manufacturing processes used are casting, turning, milling, drilling, brazing, and welding.

The main design goals for the combustion chamber and the nozzle section are nearly the same and may be listed as liner life, structural strength (and life), additional wall cooling requirements, and combustion stability devices (i.e., "acoustic cavities," used only in the combustion chamber). This leads to the definition of the hot gas wall contour, the cooling channel design, and the selection of materials and processes. The biggest design challenges are the structural and thermal design of the combustor and the nozzle extension to ensure attaining the required useful life within allowed coolant pressure losses and without exceeding the target weight for the chamber component. In the case of the nozzle extension, the superposition of thermal loads and structural loads in the form of side loads, both mainly due to cyclic loading, (engine start/stop) must be covered. If the engine operates on an expander cycle, the heating of the chamber and nozzle coolant must be maximized to drive the turbines at the required power level.

High chamber pressure results in high heat loads (e.g., 100 Btu/in.²-s. at 3000 psia chamber pressure for the SSME) and dictates the use of a high-conductivity copper alloy liner of minimum wall thickness. Today's high-pressure combustors use the milled channel design, whereas future combustors for expander cycle engines may use a tubular design. After repeated cyclic loading (engine start/stop) and high heat flux/high temperature liner operation, the liner may fail structurally, primarily due to thermal ratcheting which is largely a combination of low-cycle fatigue and thermal creep. Hence, the cooling must be designed to control the maximum temperature levels of the gas-side liner material along the coolant flow path, which allows the required chamber life. For a prescribed hot-gas flow path contour, the cooling channel or cooling tube geometry must be designed to keep the maximum liner temperature below the limit required for life, while not exceeding the available pressure drop. As a typical example, the chamber shown in Fig. 11 has 360 cooling channels and a pressure drop of about 30% of the chamber pressure. Pressure drop increases with chamber pressure, and is about 50% of the chamber pressure for the SSME.

In hydrogen/oxygen systems, a chemical effect called blanching occurs above certain liner temperatures and degrades the liner material's properties. Similar effects (chemical attack) occur with other propellant combinations, (e.g., MMH/N₂O₄) whose combustion gases are incompatible with copper alloys and require a coating to allow high-pressure operation. Lower pressures may allow using a nickel liner material that offers a compromise between liner conductivity and susceptibility to chemical attack, depending on the operating regime. In LOX/RP-1 systems, another propellant-related effect occurs below a certain pressure level. During operation, the hot gas wall is coated with a condensed carbon layer, called coating, effectively reducing the heat flux. On the coolant side, propellant material compatibility issues are hydrogen embrittlement of nickel-based alloys or coking of the coolant side wall depending on which propellant is used. These effects lead to increased coolant pressure drop and to liner material temperatures that may result in chamber burn-through.

A milled channel design requires turning and milling operations on the copper alloy (CuAgZr or CuCr), closing the channels, and attaching a structural jacket. Processes used for closing milled channel chambers include nickel electroforming optionally reinforced with a welded shell or compression brazing a high-strength jacket directly to the liner. In a different construction approach, tubular chambers are formed by brazing tubes and applying a structural jacket by one of the methods cited, or by using thermal spray techniques to apply an external metal structure. Inlet and outlet manifolds are typically inert gas welded or electron beam (EB) welded to the structural jacket.

Nozzle sections and extensions are basically formed by the same methods as described before. However, stainless steel materials usually have sufficient heat flux capability. Tubular structures are used for lightweight designs, and milled stainless steel liners may be used if weight is of minor importance. Film-cooled and radiation-cooled metallic or ceramic matrix composite nozzle sections may also be incorporated, depending on the application (e.g., the new Rocketdyne RS-68 uses a ceramic ablative nozzle).

Nontraditional Engines

Several new systems that could compete with conventional liquid rocket engines for the Earth-to-orbit mission are becoming more practical due to recent advances in technology. First and foremost among these are supersonic combustor ramjets, or scramjets. The concept of scramjets was developed some time ago, but severe technical obstacles had to be overcome before they could be considered for application. In the 1970s and 1980s, a series of breakthroughs occurred and since then several programs were begun to develop a practical vehicle using this for propulsion. The ability of scramjet engines to obtain their oxygen from the atmosphere leads to a much higher I_{sp} than possible with other liquid systems. For instance, in the 8–10 Mach number range, scramjets have an I_{sp} greater than 3000 seconds (14). This falls off as the Mach number increases. The upper speed limit of the scramjet has not been determined, but theoretically it is above the Mach 20–25 speed range required for orbital velocity. The scramjet benefit compared to a rocket becomes small as its speed approaches orbital velocity, but this

is only academic because the flight environment provides nearly insolvable structural and internal and external aerodynamic challenges.

System considerations lead to vehicles using scramjet propulsion that are very much more like airplanes than rockets. This in turn results in their ability to use aerodynamic forces, rather than rocket thrust, to control the vehicles. Besides I_{sp} , which translates to a takeoff gross weight advantage, these vehicles have advantages in flight safety (abort, fly back) and mission flexibility (launch window, orbital offset, and rapid rendezvous). The major disadvantage is technological readiness. Rockets and rocket-powered vehicles were refined and re-refined during the past 40 years. Airbreathing vehicle technology is just now coming of age.

Because the scramjet-powered vehicle obtains all of its oxygen from the inlet, the vehicle integration problem is much more severe than with a conventional rocket vehicle. The forebody of the vehicle is really a part of the inlet, and the aft end of the vehicle functions as part of the nozzle. Among other things, this places operational restrictions on scramjet-powered vehicles, such as Mach number and angle of attack limitations to ensure inlet air capture. Effective scramjet-powered vehicle design requires true synergy, where the engine and vehicle functions blur, and the only true measure of scramjet performance is the overall mission success.

Scramjets cannot operate at low speeds, so engines that are required to operate across a large speed range must include another cycle. Various combinations have been studied, and the rocket-based combined cycle (RBCC) is a popular choice because of its simplicity (i.e., use of a single flow path). RBCC engines use a rocket integrated within the scramjet combustor, as shown in Fig. 13, to provide static thrust for takeoff, subsonic, and low supersonic

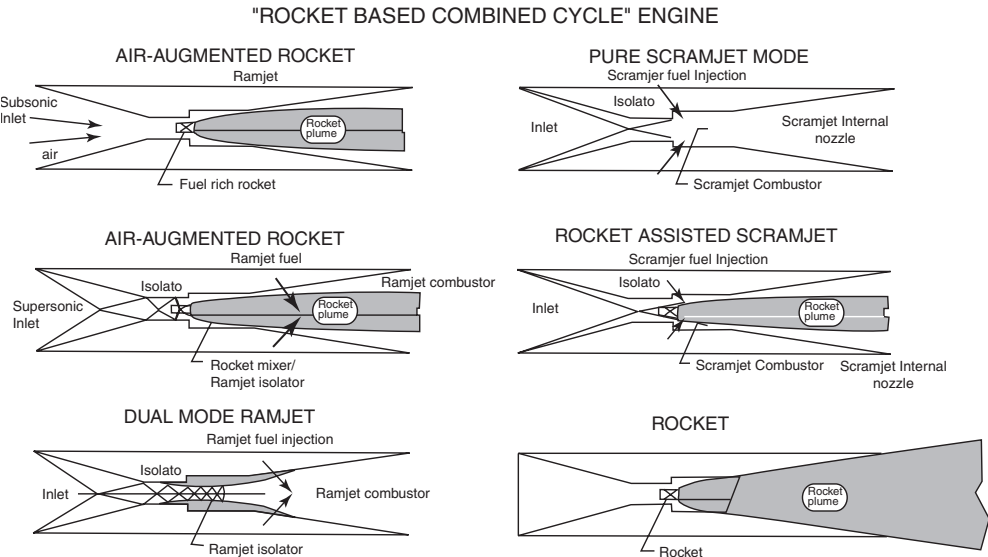


Figure 13. Rocket-based combined cycle.

operation up to ramjet speeds. Following dual-mode scramjet operation to Mach 10–16; the rocket is used first to supplement the scramjet thrust and finally alone to allow orbital insertion and in-space maneuvers. Another approach being considered uses a high-speed turbojet for flight up to Mach 4–5; then the dual-mode scramjet and separate rocket are used for higher speed. This turbine-based combined cycle has additional complexity but uses a more reliable engine cycle (turbojet). The best candidate for reduced cost, improved reliability, and safety for space access is an issue of current studies.

The scramjet engine consists of an inlet, an isolator, a combustion chamber, a nozzle, and a fuel pump and supply system. These components each have design criteria that define their configurations and performance characteristics. In addition, a scramjet propulsion system is a highly integrated aerodynamic device with strong interactions among components; each component must be designed to satisfy integration criteria for the total system.

The inlet is designed for no “spillage” in the design flight condition because spilled air is an additional drag on the aircraft. This is achieved for two-dimensional mixed compression inlets by configuring the inlet so that the first inlet shock rests on the lip of the opposing lower inlet cowl. For a typical mixed compression inlet, where the forebody of the aircraft acts as a precompression surface, the total compression is roughly split equally between the external forebody and internal to the inlet. The other primary design criterion is that the inlet provides downstream flow conditions to satisfy the requirements of temperature, pressure, and velocity necessary to achieve combustion within the confines of the combustor. Good engine performance also requires relatively uniform air distribution at the combustion inlet. Compression in scramjet inlets must be restricted so that the flow remains supersonic. For example, a Mach 7 scramjet-powered vehicle would have a combustor entrance Mach number of about 2–3. The function of the isolator is to limit the amount of pressure that is fed upstream in the wall boundary layers from the combustor, so that the inlet does not unstart. A series of oblique shocks exist in the isolator as a normal part of achieving the necessary compression.

Downstream of the inlet and isolator is the combustor. Here fuel is injected into the supersonic airflow, mixes and burns within the short time that it takes to traverse the combustor length, typically of the order of milliseconds. The location and amount of fuel injected varies with the operating and flight conditions. At low relative Mach numbers, fuel is supplied from wall injectors near the downstream end of the combustor. As speed increases, the fuel supply will gradually be moved forward until at hypersonic speeds, the fuel will be injected at the downstream end of the isolator. The design of the injector elements is extremely demanding because of the requirement for ultrarapid mixing. Variation of the fuel injection location with flight Mach number, used in combination with a diverging area combustor, provides effective throat area control and greatly simplifies combustor design issues. Because of high combustor pressure and temperature, minimization of combustor length is important to overall engine weight. Combustor cooling is also complicated by continual variation in heating patterns as shock structure changes with flight Mach number and throttle setting. A related scramjet design issue is thermal balance of the cooling and engine thrust requirement on fuel use. Thermally efficient structural design of cooling

jackets can significantly reduce cooling requirements, and thus minimum fuel use and engine specific impulse.

The end of the combustion process, where the fuel completes burning, is effectively the beginning of the nozzle. The ideal nozzle arrangement will allow the flow to expand gradually and turn along the base of the vehicle until it reaches the pressure corresponding to the atmosphere and its direction is parallel to the direction of thrust. The typical nozzle will consist of two components, the base of the vehicle itself, and a lower flap which is adjustable to correct for variations in flight conditions.

Hypersonic airbreathing engines can operate on a variety of fuels, including hydrogen and hydrocarbons. Liquid hydrogen is the choice for space-launched vehicles because of its huge heat capacity, which is used regeneratively to cool the engine and vehicle before being burned in the engine. The lower heat capacity of hydrocarbon fuels limits their application to less than Mach 8. At higher Mach numbers using hydrocarbon fuel, more fuel is required to cool the vehicle than is used for propulsion, resulting in wasted fuel.

There are a series of important technical issues facing any scramjet designer. Some of them include achieving rapid and efficient fuel mixing and combustion, effective inlet combustor isolation, limiting peak heating rates (shock interactions with the engine leading edges), balancing total cooling requirements with fuel heat capacity and engine fuel usage rate, and integrating with the vehicle airframe and low-speed systems. Finally, it must be added that the Earth-to-orbit mission is one that is so technically challenging that any advances in high temperature, lightweight materials must be used and in fact will greatly enhance mission success.

There is another class of nontraditional engines that shows great promise. These are the hybrids. These engines are composed of solid fuel grains, usually in annular form, similar to that in solid rocket motors but with no oxidizer added to the slurry before it is poured. Instead, a liquid oxidizer is used and is introduced by having it flow (in gaseous form) through the annular region. Combustion occurs at the interface. This system has many important benefits. Unlike traditional solids, thrust can be throttled or terminated simply by varying the flow of oxidizer. It is extremely easy to cast and handle because the fuel by itself can be made fairly inert. This leads to a relatively safe system. Its performance is between those of solids and liquids. On the other hand, as motor size increases, usually by lengthening the grain, efficient combustion becomes harder to achieve because of the difficulty of providing a sufficient amount of unburned oxidizer at the gas-solid boundary. To date, several small experimental motors have been successful, but no motors in the Earth-to-orbit class have been demonstrated. With improvements in understanding the physics of mixing, it is felt that the problems of hybrids will eventually be overcome.

Summary of Design Process

The overall design requirements for the propulsion system flow from the vehicle system. The vehicle system's design requirements are derived from the mission or missions that have been specified technically and programmatically. The

mission may specify that the vehicle must deliver a payload to an orbital point (i.e., LEO—low Earth orbit or GTO—geosynchronous transfer orbit). The mission could also be required to attain a Mach number at a certain altitude. All of these requirements indicate that the vehicle will have to be propelled by some type of propulsion system that must directly interface with the vehicle and will explicitly affect the propellant fraction. Additionally, the mission architecture will indicate whether the vehicle system is reusable or expendable. The flight rate will greatly affect the way the architecture is employed and the way the overall system life-cycle cost is determined. Therefore, the vehicle and mission architecture can heavily influence the development, production, and operating cost of the propulsion system as well.

Determining the type of propellant or rocket engine cycle that is optimum to maximize the mission and the cost-effectiveness of the overall vehicle system requires what is commonly known as multidisciplinary optimization or hyperfunctional integration/optimization. All of the design choices that influence the propulsion system (e.g., cycle, propellants and ratio, chamber pressure, nozzle area ratio) directly affect the design thrust size and specific impulse. They also affect the cost of the propulsion system. Those propulsion design parameters in turn influence the vehicle size and/or payload delivery capability. Thus, many propulsion and vehicle design variables must be simultaneously or interactively examined to arrive at the optimum design. Here, the primary propulsion variables that interact with the vehicle “physics” are the thrust size, engine thrust-to-weight, and the engine specific impulse or I_{sp} (vacuum and sea level, or trajectory average). Systems analysis of the integrated design problem will consider vehicle design parameters, such as vehicle thrust-to-weight, wing loading (if winged), maximum flight path thermal and structural loading (i.e., maximum dynamic pressure), and vehicle volumetric efficiency (i.e., bulk density). These vehicle design parameters will be optimized with variations on the propulsion design parameters such as chamber pressure, nozzle area ratio, and propellant ratio. If the vehicle is designed for horizontal takeoff and ascent, then the design parameters will include inlet design flow capture as well as propulsion system integration with the airframe.

Selection of the “best” vehicle and propulsion system design will be determined by which combination of design variables provides the lowest cost, highest mission performance per unit cost, and lowest complexity and risk.

BIBLIOGRAPHY

1. Humble, R.W., G.N. Henry, and W.J. Larson. *Space Propulsion Analysis and Design*, Space Technology Series—United States Dept. of Defense and NASA. McGraw-Hill, New York, 1995.
2. Huzel, D.K., and D.H. Huang. *Modern Engineering for Design of Liquid-Propellant Rocket Engines*, Volume 147, Progress In Astronautics and Aeronautics. American Institute of Aeronautics and Astronautics, Washington, DC, 1992.
3. Sutton, G.P. *Rocket Propulsion Elements—An Introduction to the Engineering of Rockets*, 6th ed. Wiley, New York, 1992.

4. Tucker, J.E. The History of the RL10 Upper-Stage Rocket Engine. In S.E. Doyle (ed.), *History of Liquid Rocket Engine Development in the United States 1955–1980*. American Astronautical Society, San Diego, 1992, p. 140.
5. Isakowitz, S.J. *International Guide to Space Launch Systems*, 2nd ed. American Institute of Astronautics and Aeronautics, Washington, DC, 1991.
6. Brooks, C.G., J.M. Grimwood, and L.S. Swenson. Chariots for Apollo: A History of Manned Lunar Spacecraft. NASA SP-4205. NASA, Washington, DC, 1979.
7. *Theoretical Performance of Rocket Propellant Combinations*. Rocketdyne Chemical and Material Technology, The Boeing Company, Rocketdyne Power and Propulsion, 1990.
8. *Theoretical Rocket Engine Propellant Summary*. Pratt & Whitney United Technologies, September 1991.
9. Gatland, K. The space pioneers. In K. Gatland (ed.), *The Illustrated Encyclopedia of Space Technology*. Orion Books, New York, 1989.
10. Chulick, M.J., L.C. Meland, F.C. Thompson, and H.W. Williams. History of the Titan liquid rocket engines. In S.E. Doyle (ed.), *History of Liquid Rocket Engine Development in the United States 1955–1980*. American Astronautical Society, San Diego, 1992.
11. McBride, B.J., and S. Gordon. Computer Program for Calculation of Complex Chemical Equilibrium Compositions and Applications. NASA Reference Publication 1311, Lewis Research Center, Cleveland, OH, 1996.
12. Stangland, M.L. *Turbopumps for Liquid Rockets Engines: Design Considerations*, Aerospace Engineering. August 1992, p. 6.
13. Balje, O.E. *Turbomachines, A Guide to Design, Selection, and Theory*. Wiley, New York, 1981.
14. Heiser, W.H., and D.T. Pratt. *Hypersonic Airbreathing Propulsion*, AIAA Education Series. AIAA, Inc., Washington, DC, 1994.

ROBERT ROSEN
NASA Ames Research Center
Moffett Field, California

MARVIN GLICKSTEIN
RUSSELL JOYNER
Pratt & Whitney
Palm Beach, Florida

ENCYCLOPEDIA OF

SPACE SCIENCE

— AND —

TECHNOLOGY

VOLUME 2

ENCYCLOPEDIA OF SPACE SCIENCE AND TECHNOLOGY

Editor

Hans Mark

The University of Texas at Austin

Associate Editors

Milton A. Silveira

Principal Engineer, Aerospace Corp.
University of Vermont

Michael Yarymovych

President International Academy of
Astronautics

Editorial Board

Vyacheslav M. Balebanov

Russian Academy of Sciences

William F. Ballhaus, Jr.

The Aerospace Corporation

Robert H. Bishop

University of Texas at Austin

Aaron Cohen

Texas A & M University

Wallace T. Fowler

University of Texas at Austin

F. Andrew Gaffney

Vanderbilt University Medical Center

Owen K. Garriott

University of Alabama

Tom Gehrels

University of Arizona at Tucson

Gerry Griffin

GDG Consulting

Milton Halem

NASA-Goddard Space Flight Center

John S. Lewis

University of Arizona at Tucson

Thomas S. Moorman

Booz Allen & Hamilton

Norman F. Ness

University of Delaware

Robert E. Smylie

National Aeronautics Space
Administration

Richard H. Truly

National Renewable Energy Laboratory

Albert D. Wheelon

Hughes Aircraft Co.

Peter G. Wilhelm

U.S. Naval Research Laboratory

Laurence R. Young

Massachusetts Institute of Technology

Alexander Zakharov

Russian Academy of Sciences

Managing Editor

Maureen Salkin

Editorial Staff

Vice President, STM Books: **Janet Bailey**

Executive Editor:

Jacqueline I. Kroschwitz

Director, Book Production and Manufacturing:

Camille P. Carter

Managing Editor: **Shirley Thomas**

Illustrations Manager: **Dean Gonzalez**

Assistant Managing Editor:

Kristen Parrish

Editorial Assistant: **Surlan Murrell**

ENCYCLOPEDIA OF

SPACE SCIENCE —AND— TECHNOLOGY

VOLUME 2

Hans Mark

Editor

Milton Silveira

Associate Editor

Michael I. Yarymovych

Associate Editor

Maureen Salkin

Managing Editor

The *Encyclopedia of Space Science and Technology* is available Online in full color
at www.interscience.wiley.com/esst

 **WILEY-INTERSCIENCE**

A John Wiley & Sons, Inc., Publication

Copyright © 2003 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permreq@wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. This advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in Publication Data:

Encyclopedia of Space Science & Technology/Hans Mark [editor].

p. cm.

Includes index.

ISBN 0-471-32408-6 (set: acid-free paper)

1. Space Science—Encyclopedias. I. Title: Encyclopedia of Space Science and Technology.

II. Mark, Hans, 1929-

QB497.E53 2003

500.5'03—dc21

2002028867

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

ENCYCLOPEDIA OF

SPACE SCIENCE

— AND —

TECHNOLOGY

VOLUME 2

M

MARS

The fourth planet from the Sun, Mars is the first planet in the solar system beyond Earth. Compared with Earth, Mars is a small planet. With a value of 1.00 for Earth, the mass of Mars is 0.107; the density, 0.719; volume, 0.149. The equatorial diameter of Mars is 4,226 miles (6,772 kilometers) as compared with that of earth of 7,960 miles (12,757 kilometers). Although Mars has been observed since early times, reliable and detailed data did not become available until commencement of the *Mariner* exploratory program in the 1960's. With the *Viking* programs of the last half of the 1970s, some important *Mariner* data had to be revised

The orbit of Mars is noticeably eccentric (0.093). The distance from Mars to the sun varies from a minimum of 129 million miles (208 million kilometers) at perihelion to a maximum of 155.3 million miles (249 million kilometers) at aphelion. Earth-bound observations of Mars are best when the planet is within a few months of opposition (when the Earth lies between the planet and the sun). During the remainder of the present century, Mars will be closest to earth on February 11, 1995, March 20, 1997, and May 1, 1999.

Mars has two satellites, or moons, Phobos and Deimos, both discovered in 1877 by Hall. Spacecraft have shown these bodies to be cratered, rocky, and chunky, and in recent years there has been serious speculation that these may not be moons in the usual sense, but rather captive asteroids. (See also Asteroids). Phobos is quite small, with dimensions of approximately 12.4×17.4 miles (20×28 kilometers) and Deimos even smaller, 6.2×9.9 miles (10×16 kilometers).

Missions to Mars. The twin *Viking* missions to Mars, each with its own lander, represented a very sophisticated and successful venture. Among some scientists there remains perplexity regarding some of the main features of the planet, notably numerous channels and rifts at one time called "canals" by

Earth-bound observers several decades ago. Knowledge of how these features look (including full-color) and their dimensions have been greatly enhanced, but the mysteries of their origins remains unknown. Earth-bound estimates and *Mariner's* measurements of Mars comparatively thin atmosphere were confirmed, a factor which detracted from the possibility of organisms living on the planet. The polar ice cap once thought to be frozen carbon dioxide has been found to be ice with possibly some frozen carbon dioxide with it. Biological experiments designed to detect living organism proved negative, but the apparently oxidizing characteristic of Martian soil has introduced new puzzles.

An interesting view of Mars taken by *Viking Orbiter 1* showing the huge Mariner Valley (Valles Marineris) is given in Fig. 1. A close-up of this extremely impressive Martian feature is given in Fig. 2.

The *Viking* missions are discussed in greater detail in a later section of this entry. Many missions preceded the *Viking* missions to Mars, and several have followed. The list below presents the chronology of missions to Mars:

Mars 1960A—USSR Mars Probe was launched on October 10, 1960, however, it failed to reach Earth orbit.

Mars 1960B—USSR Mars Probe. Launched on October 14, 1960. It also failed to reach Earth Orbit.

Mars 1962A—USSR Mars Flyby. Launched on October 24, 1962, this spacecraft failed to leave Earth orbit after the final rocket stage exploded.

Mars 1—USSR Mars Flyby. Launched on November 1, 1962. The spacecraft weighed 1,969 pounds (893 kilograms). This mission was not successful due to communications failure.

Mars 1962B—USSR Mars lander. Launched on November 4, 1962.

This spacecraft also failed to leave Earth orbit.

Mariner 3—Launched on November 5, 1964 at a weight of 572 pounds (260 kilograms) by the USA, the solar panels did not open, preventing a successful flyby. *Mariner 3* remains in solar orbit.

Mariner 4—Launched on November 28, 1964 at a weight of 572 pounds (260 kilograms) by the USA, *Mariner 4* reached Mars on July 14, 1965. It passed within 5,952 miles (9,920 kilometers) and returned data confirming that the atmosphere was composed of carbon dioxide, and identifying a small magnetic field. *Mariner 4* obtained 22 close-up photos of the surface of Mars clearly showing surface features, notably craters. *Mariner 4* remains in solar orbit. See Fig. 3.

Zond 2—USSR Mars Flyby launched on November 30, 1964, which was unsuccessful. Contact with the spacecraft was lost and its fate is unknown.

Mariner 6—USA Mars Flyby launched at a weight of 910 pounds (413 kilograms), the spacecraft reached Mars on July 31, 1969. It passed within 2,062 miles (3,437 kilometers) of the surface of the planet. *Mariner 6* remains in solar orbit.

Mariner 7—USA Mars Flyby launched at a weight of 910 pounds (413 kilograms), the spacecraft reached Mars on August 5, 1969. It passed within 2,131 miles (3,551 kilometers) of the surface of Mars at the south pole region. Both *Mariner 6* and *Mariner 7* obtained data related to the atmosphere and surface composition. Over 200 photos were obtained during these two missions. *Mariner 7* remains in solar orbit. See Fig. 4.

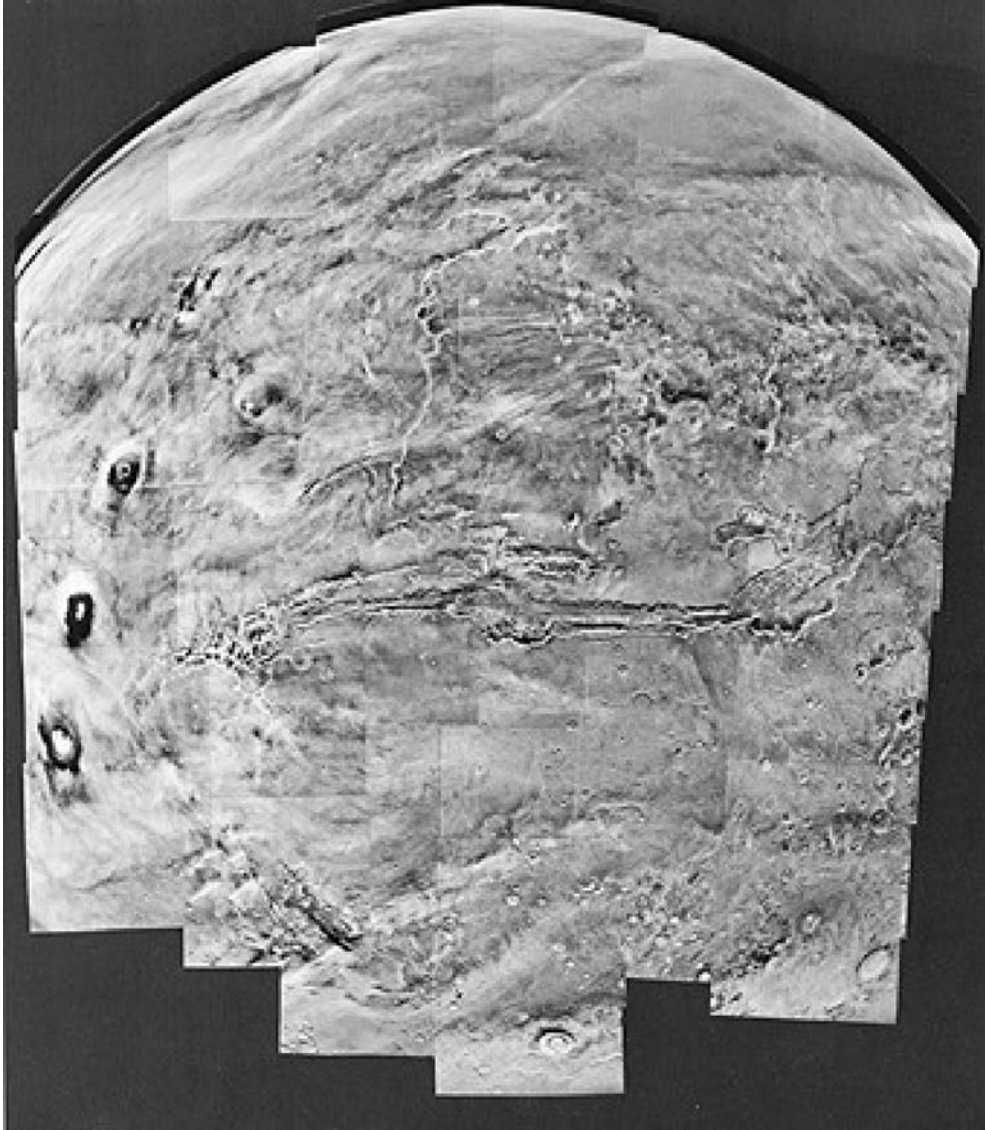


Figure 1. Mosaic of 102 photos of Mars taken on February 22, 1980 by *Viking Orbiter 1*. Several prominent Martian features and at least two unusual weather phenomena are visible. Valles Marineris (Mariner Valley), as long as the North American continent from coast to coast, stretches across the center. Three huge volcanoes of the Tharsis Ridge are visible at the left: Arsia Mons, Pavonis Mons, and Ascraeus Mons, proceeding from south to north. A sharp line, either a weather front or an atmospheric shock wave, curves north and east from Arsia Mons. This is the first time a feature like this had been seen on a planet. Four tiny clouds can be seen in the southernmost frame, just north of a large crater named Lowell. While the clouds are too close together to be resolved, even under high magnification, their shadows can be separated easily. The largest cloud is nearly 32 kilometers (20 miles) long. Measurements show the elevation of the clouds at nearly 28 kilometers (91,000 feet). Such distinct cloud-shadow patterns apparently are quite rare on Mars. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

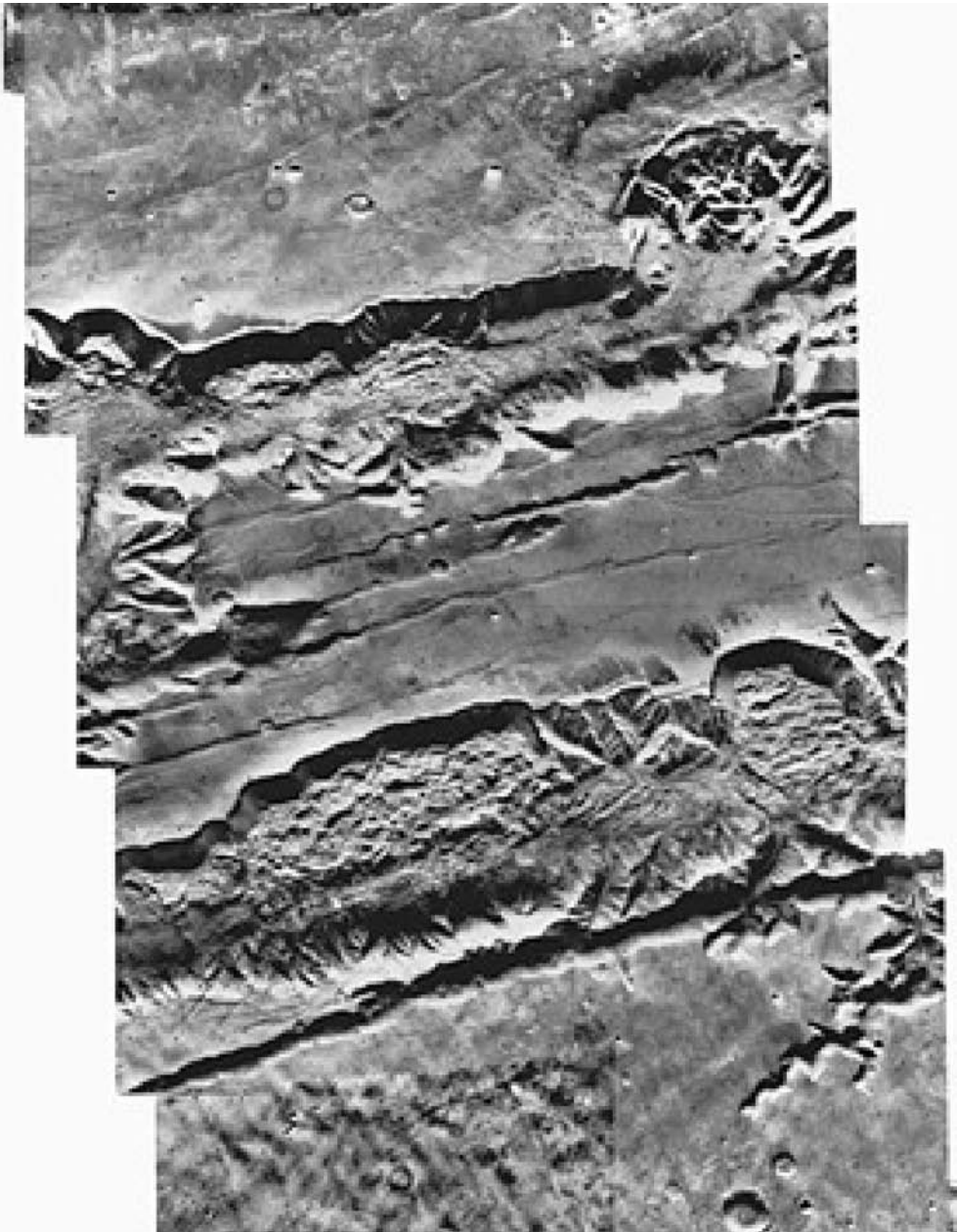


Figure 2. Mosaic of the surface of Mars showing the west end of the Valles Marineris (all of which is shown in Fig. 1) from a range of about 4300 kilometers (2700 miles). These two canyons, running east-west across the picture, are each about 60 kilometers (37 miles) wide and more than 1 kilometer (0.6 mile) deep. Some scientists suggest that the canyons were originally formed by downfaulting of the crust along parallel faults. Other faults and collapsed depressions with the same trend are seen between the two canyons. After they were formed, it is suggested that the canyons were modified by erosion that formed great slumps on the walls and also cut side valleys to the main canyons. A few comparatively recent impact craters will be noted, particularly at the bottom right of the view. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

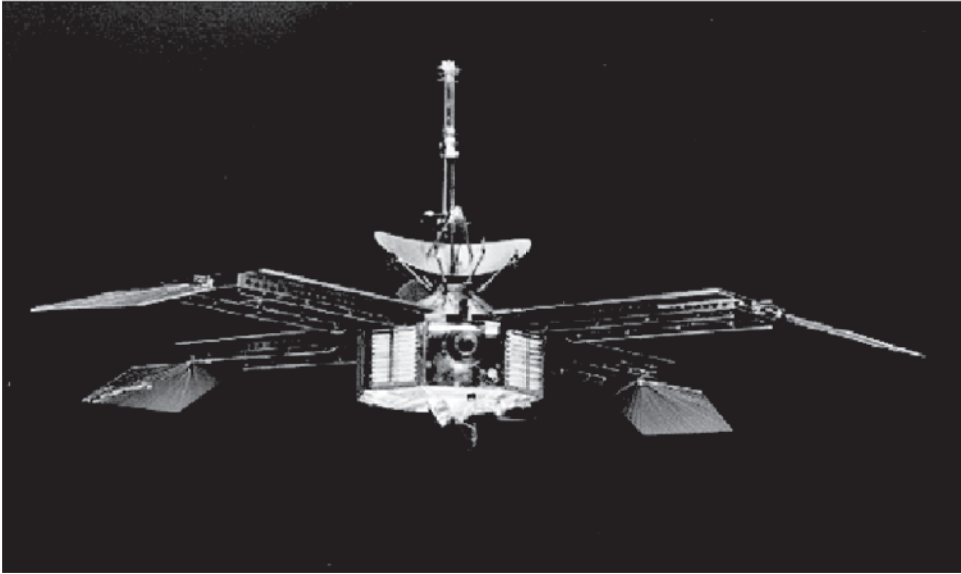


Figure 3. *Mariner 4* was the fourth in a series of spacecraft used for planetary exploration in a flyby mode. It was designed to conduct close-up scientific observations of the planet Mars and to transmit these observations to Earth. (Courtesy of the Jet Propulsion Laboratory and NASA's National Space Science Data Center.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Mariner 8—USA Mars Flyby launched May 8, 1971, this mission was unsuccessful as it failed to reach Earth's orbit.

Kosmos 419—Launched by the USSR May 10, 1971, this mission was unsuccessful as it failed to reach Earth's orbit.

Mars 2—This spacecraft was a USSR Mars Orbiter/Soft Lander launched May 19, 1971 that weighed 10,230 pounds (4,650 kilograms). It failed in its landing mission as the *Mars 2 Lander*, which was released from the Orbiter on November 27, 1971, crash-landed on the surface of the planet. It is known that the breaking rockets failed, but no data was returned. The *Mars 2 Orbiter* returned data until 1972.

Mars 3—This spacecraft was another USSR Mars Orbiter/Soft Lander that weighed 10,215 pounds (4,643 kilograms). It reached Mars on December 2, 1971, and successfully released the lander to the surface of the planet. It was the first successful landing on the surface of Mars, but the *Mars 3* failed to record and transmit more than 20 seconds of data to the orbiter. The *Mars 3 orbiter* collected data related to the surface temperature and atmospheric conditions until August 1972.

Mariner 9—Launched by the USA May 30, 1971, the spacecraft weighed 1,116 pounds (506 kilograms). *Mariner 9* was the first US spacecraft to enter orbit around a body other than the Earth's moon, and it entered this orbit on November 24, 1971. Among the data obtained were the first high-resolution images of the Martian moons, Phobos and Deimos, and surface data detailing river and channel-like features. *Mariner 9* remains in Martian orbit. See Fig. 5.



Figure 4. *Mariner 6* and *7* were designed to fly over the equator and southern hemisphere of the planet Mars. They were solar powered and capable of continuous telemetry transmission. Each spacecraft weighed 910 pounds (413 kilograms) and measured 11 feet (3.35 meters) from the scan platform to the top of the low-gain antenna. The width across the solar panels was 19 feet (5.8 meters). The eight-sided body of the spacecraft carried seven electronic compartments. A small rocket engine, used for trajectory corrections, protruded through one of the sides. The planetary experiments aboard the spacecraft were two television cameras, an infrared radiometer, and infrared spectrometer and as ultra-violet spectrometer. (Courtesy NASA.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Mars 4—Another of the USSR Mars Orbiter/Soft Lander vehicles, this mission was not wholly successful. Although it arrived at Mars in February 1974, it failed to enter orbit due to failure of the braking rockets. A flyby at a distance of 1,320 miles (2,200 kilometers) returned limited data.

Mars 5—A USSR Mars Orbiter/Soft Lander vehicle, the spacecraft weighed 10,230 pounds (4,650 kilograms) and entered Martian orbit in February 1974. Data obtained during this mission set the stage for the *Mars 6* and *Mars 7* missions.

Mars 6—This USSR Mars Orbiter/Soft Lander vehicle, which weighed 10,230 pounds (4,650 kilograms) entered Martian orbit on March 12, 1974 and launched its lander. The lander successfully transmitted atmospheric data during its descent, but failed prior to landing.

Mars 7—Another USSR Mars Orbiter/Soft Lander vehicle that weighed 10,230 pounds (4,650 kilograms), failed both to enter Martian orbit and to set the

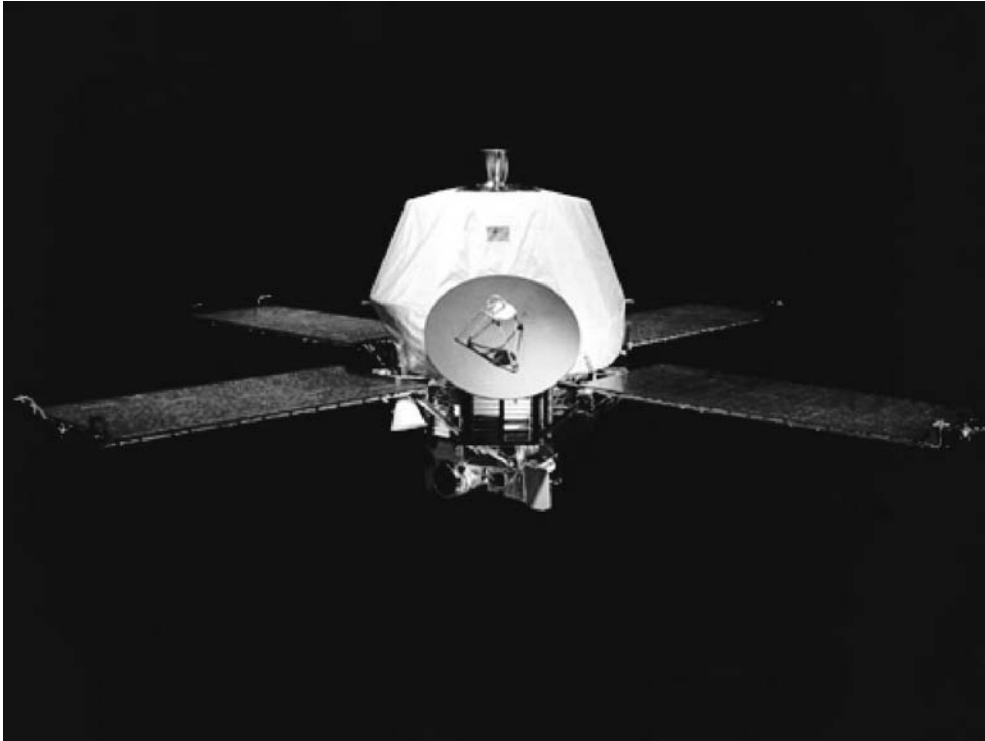


Figure 5. The *Mariner 9* spacecraft was built on octagonal magnesium frame 18 inches (45.7 centimeters) deep and 54.5 inches (138.4 centimeters) across a diagonal. Four solar panel each 85×35 inches (215×90 centimeters), extended out from the top of the frame. Each set of two solar panels spanned 23 feet (6.89 meters) from tip to tip. Also mounted on the top of the frame were two propulsion tanks, the maneuver engine, a 5-foot (1.44 meters) long low gain antenna mast and a parabolic high gain antenna. A scan platform was mounted on the bottom of the frame, on which were attached the mutually bore-sighted science instruments (wide-and narrow-angle TV cameras, infrared radiometer, ultraviolet spectrometer, and infrared interferometer spectrometer). The overall height of the spacecraft was 7.5 feet (2.28 meters). (Courtesy of NASA's National Space Science Data Center.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

lander vehicle on the Martian surface. The *Mars 7 orbiter* and lander remain in solar orbit.

Viking 1—Designed after the *Mariner* spacecraft, the USA Mars Lander/Orbiter was launched on August 20, 1975 weighing 7,478 pounds (3,399 kilograms). The orbiter weighed 1,980 pounds (900 kilograms) and the lander weighed 1,320 pounds (600 kilograms). *Viking 1* entered Martian orbit June 19, 1976, and its lander successfully set on the surface one day later on July 20, 1976 on the western slopes of Chryse Planitia. The lander and orbiter obtained data related to the weather on Mars, the Martian terrain, and microorganisms on the planet. The *Viking 1 orbiter* ran out of altitude control propellant August 7, 1980 and was deactivated. The *Viking 1 lander* was accidentally shut down and neither communication nor activation was ever regained. See Fig. 6.

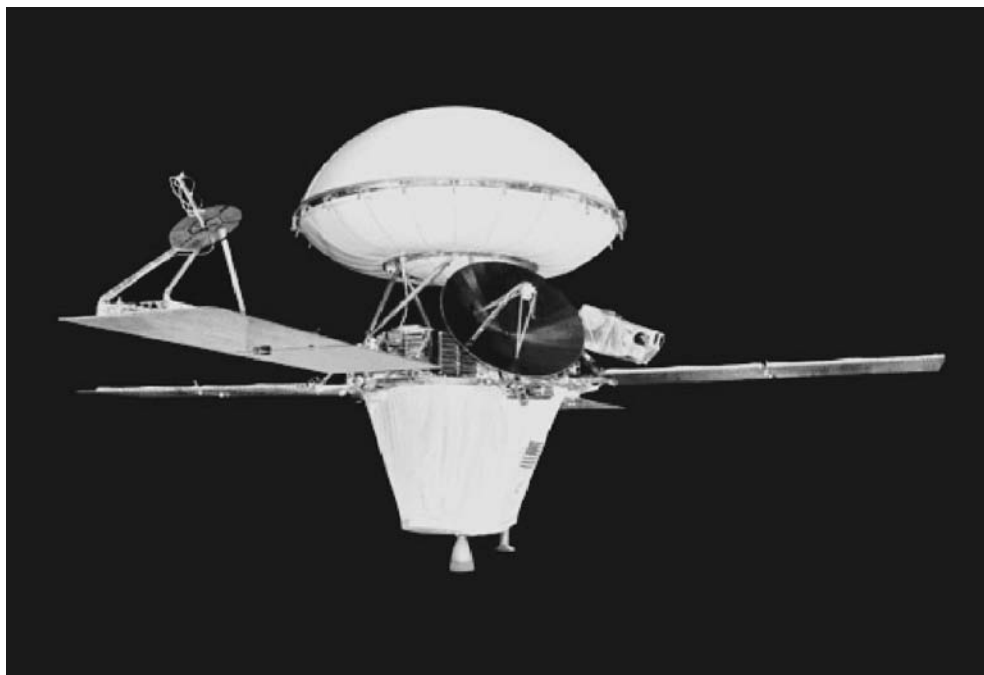


Figure 6. This image shows a model of one of the *Viking* spacecraft, which were made of two parts: an orbiter and a lander. The orbiter's initial job was to survey the planet for a suitable landing site. Later the orbiter's instruments studied the planet and its atmosphere, while the orbiter acted as a radio relay station for transmitting lander data. (Courtesy NASA/JPL.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Viking 2—Also designed after the *Mariner* spacecraft, the USA Mars Lander/Orbiter was launched on September 9, 1975 weighing 7,478 pounds (3,399 kilograms). See Fig. 7. The orbiter weighed 1,980 pounds (900 kilograms) and the lander weighed 1,320 pounds (600 kilograms). *Viking 2* entered Martian orbit on July 24, 1976 and its lander set down at Utopia Planitia on August 7, 1976. See Fig. 8. While both landers had experiments to search for and identify microorganisms on Mars, the results of the experiments are still subject to debate. Both landers together obtained over 52,000 images while mapping the planet's surface. The *Viking 2 orbiter* ran out of altitude control propellant July 25, 1978 and was deactivated. Because the *Viking 2 lander* used the *Viking 1 orbiter* for communications, it had to be shut down the same time the *Viking 1 orbiter* was deactivated on August 7, 1980.

Phobos 1—USSR Mars Obiter/Lander weighing 11,000 pounds (5,000 kilograms) that was launched on July 7, 1988. The spacecraft was lost on the way to Mars due to a command error on September 2, 1988. See Fig. 9.

Phobos 2—USSR Phobos Flyby/Lander, which weighed 11,000 pounds (5,000 kilograms), was launched on July 12, 1988. The spacecraft entered Martian orbit January 30, 1989, but failed at a distance of 480 miles (800 kilometers) from the Martian moon Phobos.



Figure 7. Launch of the *Viking 2* spacecraft from Cape Canaveral, Florida. (Courtesy NASA/JPL.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Mars Observer—USA Mars Orbiter was launched September 25, 1992 but failed to enter Martian orbit. Communication with the Mars Observer was lost August 21, 1993.

Mars Global Surveyor—USA Mars Orbiter was launched November 7, 1996 to complete the mission of the Mars Observer. See Fig. 10.

Mars 96—Russia Orbiter and Lander which was launched November 16, 1996, consisted of an orbiter, two landers and two soil penetrators. The fourth stage of the rocket that launched the Mars 96 spacecraft ignited prematurely as the vehicle entered orbit. The spacecraft crashed into the ocean and sank between the coast of Chile and Easter Islands.

Mars Pathfinder—USA Lander and Surface Rover launched on December 4, 1996. The lander weighed 581 pounds (264 kilograms) and the rover vehicle weighed only 23 pounds (10.5 kilograms). Mars Pathfinder reached Mars July 4, 1997 and impacted the surface at a velocity of approximately 40 miles per hour (18 meters per second).



Figure 8. Captured in this rendering is a *Viking* lander just before it touched down on the Martian surface. The parachute and upper aeroshell can be seen in the upper left corner of the image. At this stage of the descent, the lander's terminal descent propulsion system (three retro-engines) had slowed the craft down so that velocity at landing was about 7 miles per hour (2 meters per second). Seconds after the lander reached the surface it began transmitting images back to the orbiter for relay to Earth. (Courtesy NASA/JPL.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

It bounced into the air about 50 feet (15 meters), bounced another 15 times, and rolled to a stop approximately 2.5 minutes after impact about one-half mile (about 1 kilometer) from the site of initial impact. The landing site, in the Ares Valley region at 19.33°N, 33.55°W, was named the Sagan Memorial Station. The rover, a six-wheeled vehicle named Sojourner, hit the Martian surface July 6. See Figs. 11 and 12. This highly successful mission returned 2.6 billion bits of information including over 16,000 images from the lander, 550 images from Sojourner, 15 chemical analyses of rocks, and extensive data on climatic conditions. (See also Pathfinder Mission to Mars).

Nozomi. Launched on July 3, 1998, by Japan's Institute of Space and Astronautical Sciences, it will be the first Japanese spacecraft to reach another planet. Nozomi will study Martian aeronomy, particularly Mars' upper atmosphere, ionosphere, and their interactions with the solar wind. The spacecraft will reach Mars and go into orbit in December 2003.



Figure 9. This artist's concept depicts the *Phobos 1 & 2* spacecraft destined for Mars. They were the next-generation in the Venera-type planetary missions, succeeding those last used during the Vega 1 and 2 missions to comet P/Halley. (Courtesy of NASA's National Space Science Data Center.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Mars Climate Orbiter. Launched on December 11, 1998, to embark on a study of the planet's climate, daily weather and current water resources. On September 29, 1999, the spacecraft was lost due to a navigational error as it entered the atmosphere of Mars (See Figure 13).

Mars Polar Lander. Launched on January 3, 1999, and targeted for a landing zone near the edge of the South Polar-layered terrain. The spacecraft was designed to search for near surface ice and possible records on the surface of cyclic climate and geophysical changes. The spacecraft was lost on arrival at Mars on December 3, 1999 (See Figure 14).

NASA Unveils Its 21st Century Mars Campaigns

The seven-month retooling of its Mars campaign was prompted by the back-to-back loss last year of two spacecraft at the Red Planet. Subsequent investigative

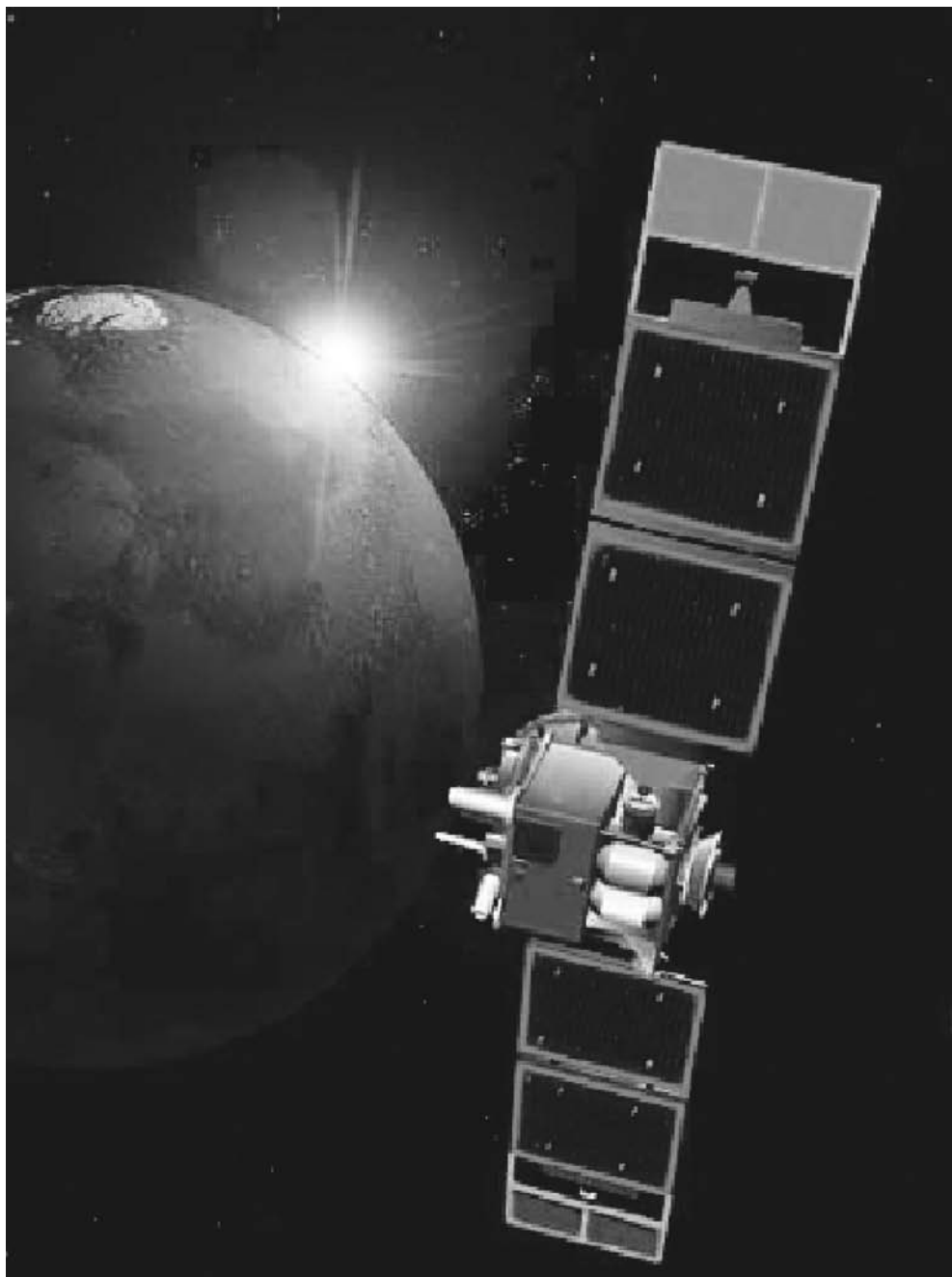


Figure 10. Captured in this rendering the *Mars Global Surveyor (MGS)* is designed to orbit Mars over a two year period and collect data on the surface morphology, topography, composition, gravity, atmospheric dynamics, and magnetic field. These data will be used to investigate the surface processes, geology, distribution of material, internal properties, evolution of the magnetic field, and the weather and climate of Mars. (Courtesy of NSSDC.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



Figure 11. Artist view of *Pathfinder* on Mars. (Courtesy of NASA's National Space Science Data Center.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

reports, including one authored by retired Lockheed Martin executive Tom Young, found bad management, a lack of training and an inadequate system of checks and balances, as well as too-tight budgets, doomed the Mars Climate Orbiter and Polar Lander missions, a \$300 million-plus loss.

NASA halted ambitious plans to send a lander/orbiter pair to Mars every 26 months, when the Earth and the Red Planet are closely aligned. Instead, it will now stagger the pace dispatching just one of each at the roughly two-year intervals.

The revised program also looks out beyond returning a sample of Martian soil to Earth for study. That goal has been pushed back to 2011 or later.

This program will represent a long-term strategy. It won't just end with Mars sample return like the previous one did. Officials, said the new program, allows for NASA to respond to any new discoveries on Mars, like the evidence

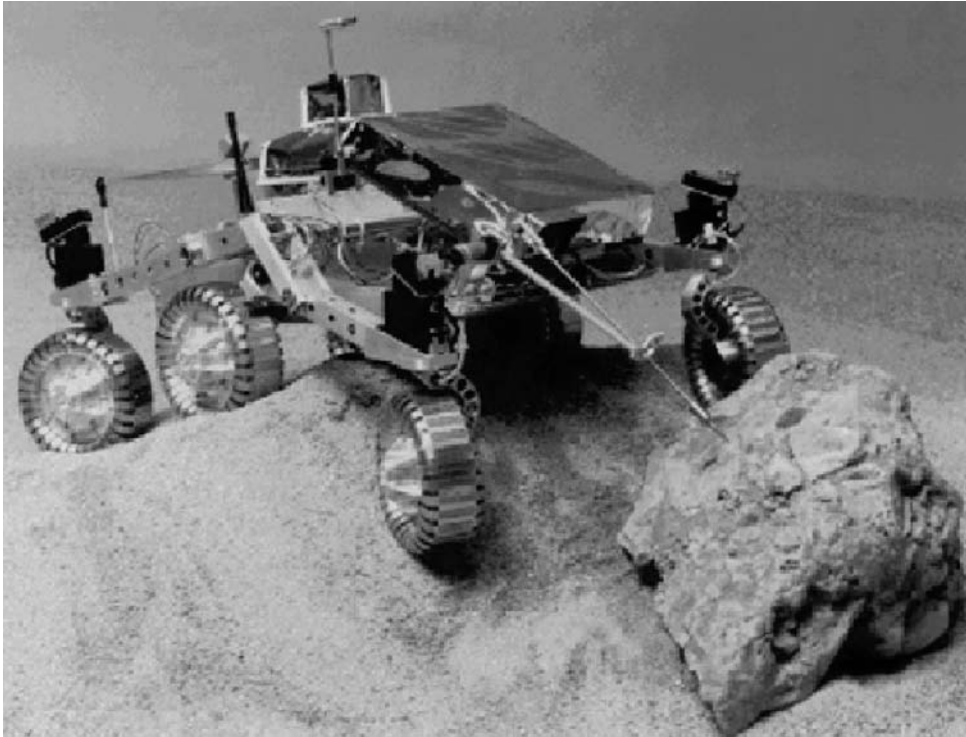


Figure 12. The rover *Sojourner* is a six-wheeled vehicle launched with the Mars Pathfinder mission. It is controlled by an Earth-based operator, who uses images obtained by both the rover and lander systems. Note that the time delay is about 10 minutes, requiring some autonomous control by the rover. The primary objectives were scheduled for the first seven sols (1 sol = 1 martian day = ~ 24.7). (Courtesy of NASA's National Space Science Data Center.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

that suggests water may have flowed on the planet's surface in the recent past, as well as accommodate the prospect of any of the planned missions failing.

What's missing from the equation are humans. NASA has already scrapped plans to launch in 2001, a package of experiments that would have laid some of the groundwork for future human missions to Mars, including experiments to produce oxygen from the Martian atmosphere and to assess the threat of its dust and radiation environments. Now, similar experiments might not make to Mars until 2007 at the earliest.

The agency plans on six major Mars missions for this decade alone, spending as much as \$450 million a year on its near-term efforts. The missions include:

2001 — Mars Surveyor Lander 2001. This program was canceled as part of the reorganization of the NASA Mars exploration program described previously. However, elements of the original program will be used on future missions. The orbiter, renamed *Mars Odyssey*, will nominally orbit Mars for three years, with the objective of conducting a detailed mineralogical analysis of the planet's surface from orbit and measuring the radiation environment. The mission has as its

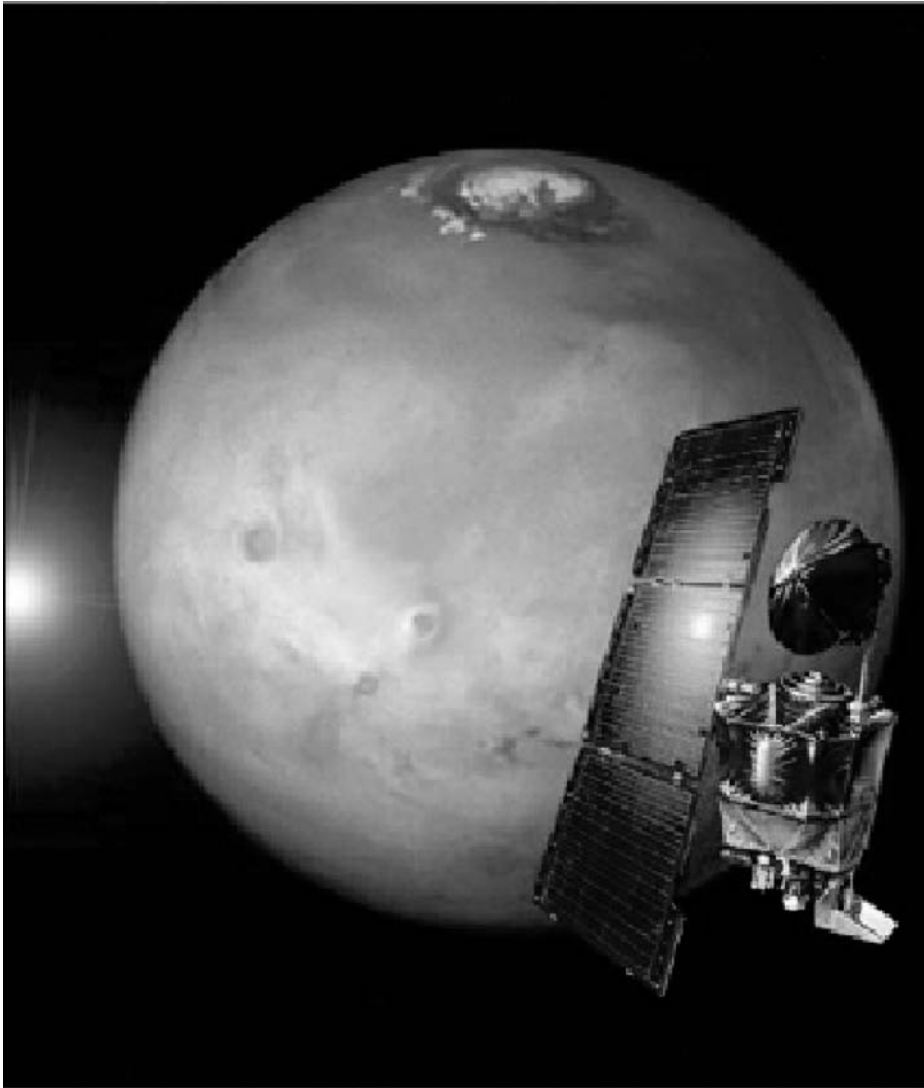


Figure 13. The *Mars Climate Orbiter* Spacecraft was launched from Cape Canaveral on December 11, 1998. It was lost on September 29, 1999, due to a navigational error. The intention was to place the spacecraft in a highly elliptical orbit around Mars and then to lower the periapsis into the upper atmosphere of Mars. The spacecraft would then use aerobraking to circularize the orbit before initiating science operations. Using this maneuver makes it possible to save substantial fuel and therefore spacecraft weight. (Courtesy NASA/JPL.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

primary science goals to gather data to help determine whether the environment on Mars was ever conducive to life, to characterize the climate and geology of Mars, and to study potential radiation hazards to possible future astronaut missions. The orbiter will also act as a communications relay for future missions to



Figure 14. The *Mars Polar Lander* Spacecraft was launched from Cape Canaveral on January 3, 1999. The spacecraft was lost on December 3, 1999, during Mars orbit insertion. The payload included instruments for sampling soil at the surface and also within one meter of the surface. In addition, there was a stereoscopic camera and also meteorological sampling equipment. (Courtesy NASA/JPL.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Mars over a period of five years. Eventually, a follow-on mission with a Lander named Marie Curie will also be flown. (See Figures 15, 16 and 17.)

2003—Two water-sniffing Mars Exploration Rovers, for detailed field geological work.

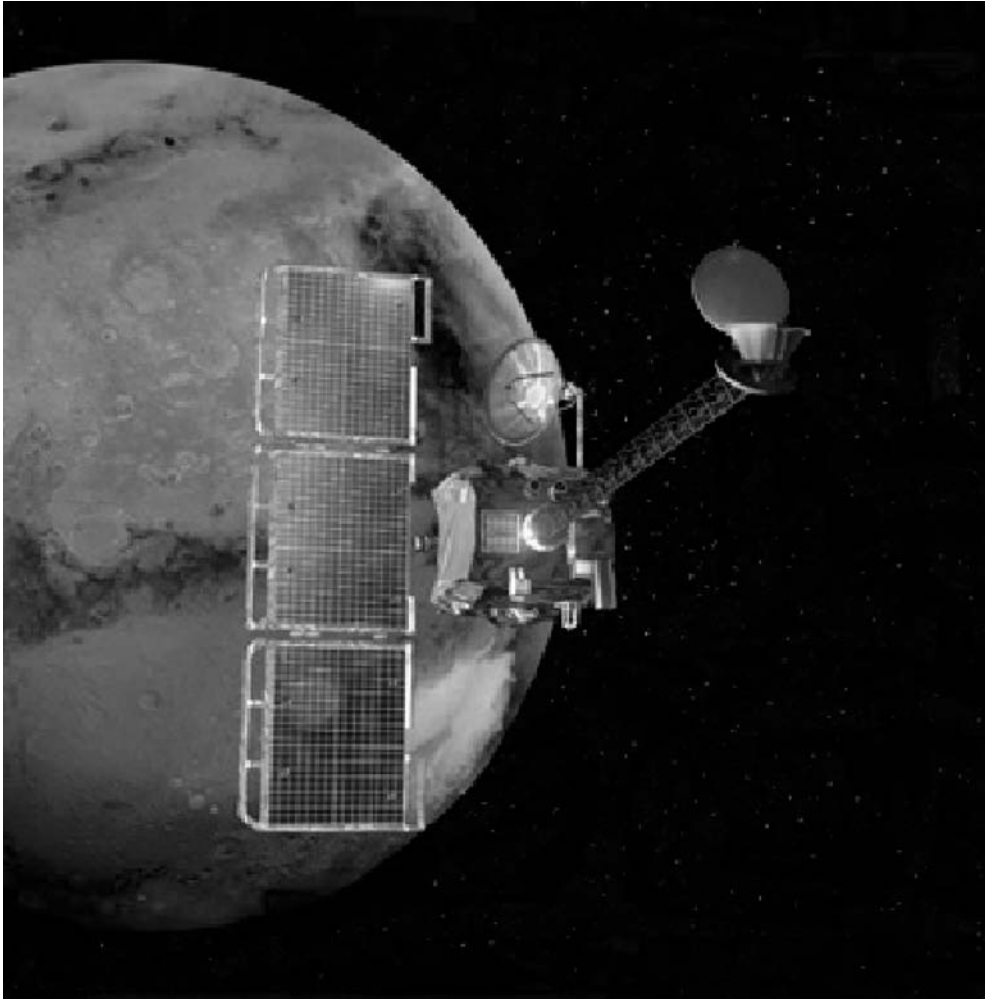


Figure 15. *The Mars Odyssey* spacecraft was originally part of the Mars Surveyor Lander 2001 plan that was canceled. The Odyssey will be the first to use the atmosphere of Mars to slow down and directly capture a spacecraft into orbit in one step, using a technique called aerocapture. It will then reach a circular mapping orbit within about 1 week after arrival. The Odyssey will carry 2 main science instruments, the Thermal Emission Imaging System (THEMIS) and the Gamma Ray Spectrometer (GRS). THEMIS will map the mineralogy and morphology of the Martian surface using a high resolution camera and a thermal infrared imaging spectrometer. The GRS will achieve global mapping of the elemental composition of the surface and the abundance of hydrogen in the shallow sub-surface. The gamma ray spectrometer was inherited from the lost Mars Observer mission. (Courtesy of NASA/JPL, 2001 Artwork by Corby J. Waste.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

2005—A Mars Reconnaissance Orbiter: an orbiter modeled on NASA's successful Mars Global Surveyor, but capable of imaging objects as small as 30 centimeters (a foot) in diameter. Jim Garvin, NASA's Mars exploration program scientist, likened it to putting a microscope in orbit.



Figure 16. The *Mars Surveyor Lander 2001*. This spacecraft was part of the canceled program already mentioned in the caption for Figure 15. It is likely to be used eventually in a future NASA Mars exploration flight. The Lander will carry an imager to take pictures of the surrounding terrain during its' rocket-assisted descent to the surface. The descent imaging camera will provide images of the landing site for geologic analyses, and will aid planning for initial operations and traverses by the rover. The Lander will also be a platform for instruments and technology experiments designed to provide key insights to decisions regarding successful and cost-effective human missions to Mars. Hardware on the Lander will be used for an in-situ demonstration test of rocket propellant production using gases in the Martian atmosphere. Other equipment will characterize the Martian soil proper and surface radiation environment. (Courtesy of NASA/JPL, 2001 Artwork by Corby J. Waste.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

2007—A “smart” surface lander equipped with a hazard avoidance system, precision landing capability and designed to deliver a rover laden with up to 270 kilograms (600 pounds) of scientific instruments; also in 2007, a “Scout” mission, which could entail a small Beagle 2-type lander, a balloon or an airplane. Both balloon and airplane Mars missions have been submitted as proposals in the current round of Discovery-class NASA projects.

The Mars Pathfinder landing in 1997 was within a 100×300 -kilometer (60×200 -mile) landing ellipse. “Where we want to be by 2007 is down to some-

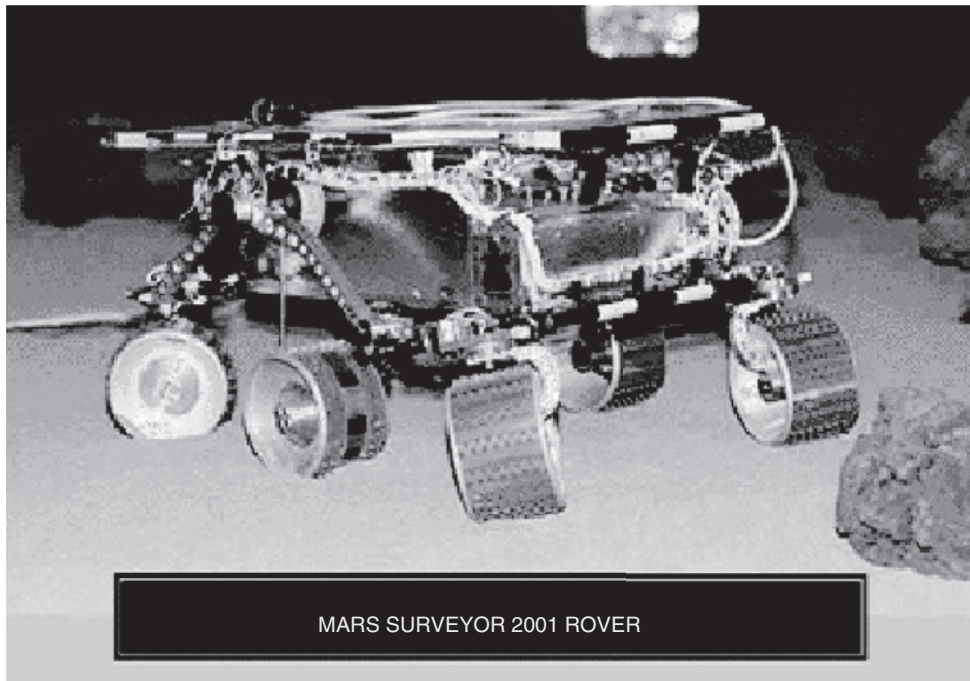


Figure 17. As part of the Mars exploration mission redesign, the plans for this rover have changed significantly. Current plans are to send the Marie Curie rover to Mars on a future lander. This rover is very similar to the Pathfinder Sojourner Rover, and in fact is the same rover that was used within the Pathfinder “sandbox” test bed pictured above. (Courtesy of NASA/JPL.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

thing that’s 1 kilometer by 3 kilometers (0.62 by 2 miles)—a reduction by a factor of 100,” Lavery said. The eventual goal is to land spacecraft on the equivalent of a Martian dime—within a tight ellipse just a few hundreds of yards (meters) across, he said.

2007—NASA could also kick off an international collaborative effort in 2007, teaming up with the Italian space agency on a telecommunications orbiter for Mars or with the French on a network of small landers.

2009—NASA could team up again with the Italians on a follow-on to the European Space Agency’s 2003 Mars Express mission. The probe would carry ground-penetrating radar to prospect for water on the Red Planet.

2011—As early as 2011, but perhaps slipping to 2014, NASA could start a long-term project to return multiple samples of Martian soil and rock to Earth. The effort, which could cost as much as \$2 billion a pop, had been on tap for 2005 under the previous plan.

The Martian Atmosphere

As pointed out by Anders and Owen (1977), the thinness of the Martian atmosphere has been one of the great disappointments of the space age. At one time,

Lowell, Dollfus, and others had suggested a surface pressure near 80 millibars (1 standard atmosphere on earth = 1013 millibars). Even with an atmosphere only about one-tenth the value of earth's, it was envisioned that Mars possibly could sustain some forms of life. *Viking 1* established a figure of 7.65 millibars. Substantiation of this relatively low pressure, indicating that Mars has only about 3% of the volatiles found on earth, revised scientific thinking in terms of the development of Mars. Anders and Owen suggested five processes, in combination, which may have been responsible for the thin martian atmosphere: (1) a small initial endowment of volatiles; (2) incomplete outgassing from the interior; (3) recondensation or trapping in surface regions; (4) catastrophic loss of an early atmosphere; and (5) gradual escape of the lighter constituents. Although these investigators do not describe a detailed model for how the Martian atmosphere passed from an early, dense state to its present condition, they offer a schematic scenario:

The basic notion is that the atmosphere gradually decreased in density as a result of the deposition of carbon dioxide in the form of carbonates and the escape of nitrogen from the upper atmosphere. While the latter process was critical for the ultimate nitrogen abundance and isotope ratio, it should have played a small role in determining the total atmospheric pressure, since carbon dioxide was probably the most abundant gas. The depositional process (which may have included formation of nitrates or nitrites) was most active during the time when liquid water was most abundant—the cutting of the sinuous channels was thus a premonition of the end of the dense atmosphere. The apparent absence of an active Martian biota has prevented the recycling of volatiles through biological processes. Moreover, there is evidence that carbonates may form even under the present arid conditions on Mars.

Anders and Owen proceed in their interesting paper to make comparisons between the large and the small planets, with earth and Mars as paradigms.

Water and Water Vapor. An infrared spectrometer operating at the 1.38. micrometer region, mounted on the scan platform, was used to detect water vapor in the Martian atmosphere.¹ This scanning device was used to measure the latitudinal variations and diurnal variations. By operating over a complete martian year, the instrument was able to measure the seasonal changes. The southern hemisphere, which was at the onset of winter, was found to have very little water vapor (0 to 0.3 precipitable micrometer). In contrast, the northern hemisphere showed a significant amount of water (up to 75 micrometers at 70–80 °N), a range of almost three orders of magnitude. The north polar region showed a slight drop in water vapor abundance. A strong diurnal repetitive cycle in certain regions, peaking out in the local mid-afternoon, was also found. Negative correlation existed between the elevation and the water vapor abundance, as would be expected. On the basis of the abundance of water vapor in the polar region, a lower limit can be put on the atmospheric temperature, namely 205 K (–68 °C; –90 °F). This value indicates that the permanent polar cap consists of water ice, a factor that was confirmed by the infrared thermal mapper (IRTM) also part of the *Viking* instrumentation package.

Observations of the latitude dependence of water vapor made from the *Viking 2 Orbiter* showed peak abundances in the latitude band 70–80 °N in the

¹Information extracted from official NASA Langley Research Center report.

northern midsummer season. Total column abundances in the polar regions appeared to require near-surface atmospheric temperatures in excess of 200 K (-73°C ; -99°F) and are incompatible with the survival of a frozen carbon dioxide cap at martian pressures. The remnant or residual on the north polar cap and the outlying patches of ice at lower latitudes are thus believed to be predominantly water ice whose thickness can be established within widely spaced limits, between 1 meter and 1 kilometer (~ 3 and 3280 feet). Broadband thermal and reflectance observations of the Martian north polar region in late summer yielded temperatures for the residual polar cap near 205 K (-68°C ; -90°F). Residual cap and several outlying smaller deposits appeared to be water ice with included dirt. No evidence was found for a permanent carbon dioxide polar cap.

The first evidence of the direct visible exchange of water between the martian surface and atmosphere was obtained by the *Viking 1 Orbiter* on July 24, 1976, as shown by Fig. 18.

Ground Ice on Mars. As reported by Squyres (NASA Ames Research Center) and Carr (U.S. Geological Survey), many martian landforms suggest the former presence of ground ice or water, including fretted and chaotic terrain, valley systems, outflow channels, and, with less certainty, various types of patterned

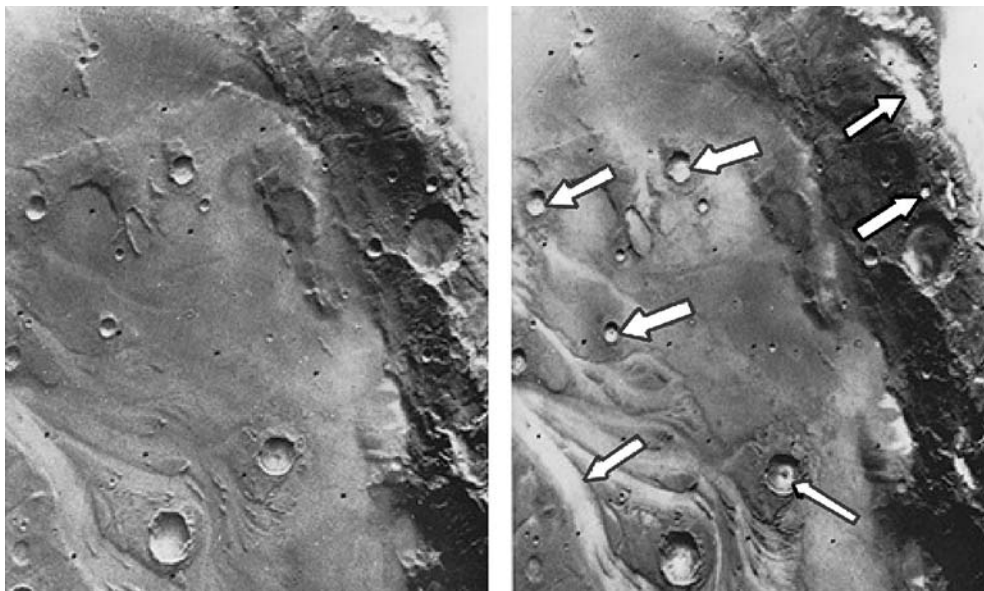


Figure 18. Two pictures taken a half-hour apart by *Viking 1 Orbiter* shows the development on Mars of early morning fog in low spots, such as crater and channel bottoms (see arrows on view at right). The scene at left was photographed shortly after martian dawn on July 24, 1976 from 12,400 kilometers (7700 miles) and, at right, 30 minutes later from 9800 kilometers (6100 miles). Slight warming of the sub-zero surface by the rising sun evidently drove off a small amount of water vapor which recondensed in the colder air just above the surface. Brightness measurements of the resulting fog patches indicated that a film of water about one micrometer thick had condensed. These fog patches were the first direct, visible evidence as to where the exchange of water between the martian surface and atmosphere may occur. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

ground and rampart craters. If sufficient ice is present now, the regolith should undergo quasi-viscous flow due to creep deformation of the ice. Accordingly, to determine where ice may be present, the researchers examined approximately 24,000 *Viking Orbiter* images taken within 5000 km of the surface and mapped the distribution of three types of features—lobate debris aprons, concentric crater fill, and terrain softening—that may indicate creep of near-surface materials.² In summary, the researchers observe that the origin of ice in the martian regolith is unclear at this time. The many lines of evidence implying that ice was common in the cratered uplands early in Martian history suggest that the ice was emplaced during an early period of intense outgassing. An alternative scenario would be the continuous outgassing throughout the planet's history at a rate substantially lower than the low-latitude depletion rates in order to keep the low latitudes ice-free. In either case, intense early meteoritic brecciation was probably largely responsible for the apparent capability of the deep regolith to hold large amounts of water.

Juvenile Water from Volcanism. In studying *Viking* data, Greeley (Arizona State University) reports that volcanism played a dominant role in the evolution of the Martian surface and environment. It is estimated that volcanism has occurred from at least the close of the period of heavy impact cratering (~3.9 billion years ago) to the age of the youngest rocks visible on the planet. Materials that are considered to be of volcanic origin cover more than half the surface of Mars. Perhaps the assumption can be made that, as with the Earth, juvenile water³ was released on Mars in association with the eruption of these volcanic materials. By determining the volumes and ages of volcanic units and inferring the volatile content for the magmas, the amounts and timing of associated water release can be estimated. Tentative conclusions indicate the amount of juvenile water release on Mars would equal a layer some 46 meters deep over the entire planet. Most of this water was released in the first 2 billion years of Martian history. There are several uncertainties in estimates made to date, including lack of knowledge of volatile contents for magmas; even terrestrial values, as used, have large uncertainties, and extrapolation to martian values is difficult. Uncertainties that stem from estimates of volcanic unit volumes can be reduced through more detailed mapping and determination of flow thicknesses, which will include additional new data, obtained from the *Mars Observer* in the 1990s.

Carbon Dioxide. At both *Viking* landing sites, it was found that the temperature was appreciably above the carbon dioxide condensation boundary, thus precluding the occurrence of carbon dioxide hazes in northern summer at

²*Lobate debris aprons*—accumulations of erosion debris at the base of steep escarpments. *Concentric crater fill*—develops where debris aprons are confined within impact craters, and inward flow of material produces a pattern of concentric ridges. *Terrain softening*—a distinctive style of landform degradation apparent in high-resolution images. Definitions as given by Squyres and Carr.

³In terms of Earth, *juvenile water* refers to water which has been derived from the crystal rocks or from the interior of the Earth, and at the time of its appearance in the circulating water of the hydrosphere represents an accretion to the available water supply. Juvenile water is, therefore, water which has not previously been a part of the hydrosphere. Although many investigators have tried to devise means for identifying juvenile water, no satisfactory method has been found. It is difficult to be certain that a particular water, such as that charged by hot springs, geysers, fumaroles, etc., is of juvenile origin. Magmatic water is derived from a magma or included in a magma (rock melt).

latitudes at least 50°N . Thus, the ground level mists seen in these latitudes would appear to be condensed water vapor. Neutral mass spectrometers carried on the aeroshells of both *Viking* spacecraft indicated that carbon dioxide is the major constituent of the martian atmosphere over the height range of 120 to 200 kilometers (74 to 124 miles). Densities for carbon dioxide measured by the upper atmospheric mass spectrometers on both *Viking* spacecraft were analyzed to yield height profiles for the temperature of the Martian atmosphere between the aforementioned range. The upper atmosphere of Mars was found to be surprisingly cold with an average temperature of less than 200 K (-73°C ; -99°F). The atmosphere contains detectable concentrations of nitrogen, argon, carbon dioxide, molecular oxygen, atomic oxygen, and nitric oxide. The upper atmosphere exhibits a complex and variable thermal structure and is well mixed to heights in excess of 120 kilometers (74 miles).

Carbon Dioxide and Heat Balance at Polar Caps. As reported by Paige and Ingersoll (California Institute of Technology), the Infrared Thermal Mappers (IRTM) aboard the two *Viking* orbiters obtained solar reflectance and IR emission measurements of the martian north and south polar regions during an entire martian year. The observations were used to determine annual radiation budgets and to infer annual carbon dioxide frost budgets. The results provide further confirmation of the presence of permanent CO_2 frost deposits near the south pole and show that the stability of these deposits can be explained by their high reflectivities. In the north, the observed absence of solid CO_2 during summer was primarily the result of enhanced CO_2 sublimation rates due to lower frost reflectivities during spring. The results suggest that the present asymmetric behavior of CO_2 frost at the Martian poles is caused by preferential contamination of the north seasonal polar cap by atmospheric dust. The investigators emphasize that the *Viking* results have made it clear that the annual heat balance at the Martian poles is not purely a local phenomenon, but may be strongly influenced by the complex, global scale geologic and atmospheric processes that bring dust to the polar regions each year. The forthcoming *Mars Observer* mission, which will follow a polar orbit about the planet, will furnish much needed additional data.

Nitrogen. Results from the neutral mass spectrometer carried on the aeroshell of *Viking 1* spacecraft showed evidence for NO in the upper atmosphere of Mars and indicated that the isotopic composition of carbon and oxygen is similar to that of earth. Mars is enriched in ^{15}N relative to Earth by about 75%, a consequence of escape that implies an initial abundance of nitrogen equivalent to a partial pressure of at least 2 millibars. The initial abundance of oxygen present either as carbon dioxide or water must be equivalent to an exchangeable atmospheric pressure of at least 2 bars in order to inhibit escape-related enrichment of ^{18}O . McElroy, Yung, and Nier (1976) constructed models for the past history of nitrogen on Mars based upon *Viking* measurements showing that the atmosphere is enriched in ^{15}N . The enrichment is attributed to selective escape, with fast atoms formed in the exosphere by electron impact dissociation of N_2 and by dissociative recombination of N_2^+ . The initial partial pressure of N_2 should have been at least as large as several millibars and could have been as large as 30 millibars if surface processes were to represent an important sink for atmospheric HNO_2 and HNO_3 .

Krypton and Xenon. These gases were discovered in the Martian atmosphere with the mass spectrometer on *Viking Lander 2*. Krypton is more abundant than xenon. The relative abundances of the krypton isotopes appear normal, but the ratio of xenon-129 to xenon-132 is enhanced on Mars relative to the Earth value for this ratio. The mass spectrometer on *Viking Lander 1* had previously reported the detection of ^{36}Ar and the establishment of upper limits on Ne, Kr, and Xe in the atmosphere. The upper limit of krypton was close the value that would be predicted if the $^{36}\text{Ar}/\text{Kr}$ ratio on Mars were identical to that on Earth. As pointed out by Owen et al. (1976), the Earth's atmosphere is deficient in xenon compared with the primordial gas in meteorites, and this is exactly the situation found on Mars. The xenon deficiency on Earth has been attributed to the preferential adsorption of xenon in shales and other sedimentary material after it was out-gassed. One is thus led to the tentative conclusion that similar processes have been active on Mars, perhaps in association with the epochs of fluvial erosion that have left their imprint on the planet's surface. Owen et al. suggest as an alternative or supplementary suggestion that some of the xenon could be absorbed in the regolith.

Weather

The atmosphere of Mars appears to favor stability, much unlike Earth in that regard. The annual temperature range for the Martian surface at the *Viking* landing sites was computed on the basis of thermal parameters derived from observations made with infrared thermal mappers. *Viking Lander 1* site showed small annual variations in temperature, whereas VL-2 site showed larger annual changes. (Locations of the sites are described in the latter portion of this article.) At both sites, daily temperature ranges at the top of the soil were 183 to 268 K (-90 to -5°C ; -130 to $+23^\circ\text{F}$). Diurnal variations decreased with depth in an exponential manner. The maximum temperature of soil sampled beneath rocks at the VL-2 site was computed to be 230 K (-43°C ; -45°F). Daily patterns of temperature, wind, and pressure were highly repetitive at both sites during the early summer period. Wind was found to have a vector mean of 0.7 meter/second from the southeast with diurnal amplitude of 3 meters/second. Pressure exhibited both diurnal and semidiurnal oscillations, although of substantially smaller amplitude than those of VL-1. It should be mentioned that Mars does not have an ozone layer in its atmosphere as a shield against ultraviolet radiation and the absence of it, of course, has some effect on its climate.

It will be obvious to the reader that a satisfactory model of Martian weather cannot be formulated with so many unanswered questions as thus far indicated in this article. One popular concept, based upon incomplete data, suggests that the Earth and Mars commenced largely with similar initial conditions, but that over the course of some 4 billion years, the two planets have evolved differently. Both of these premises seem reasonable and logical. The images from *Viking* certainly prove that the two planets are distinctly different, yet when compared with other planets in the solar system. Earth and Mars present more similarities. Fundamental differences (not a direct function of the passage of time) between the two planets include: (1) *size* (Mars is only a little more than half the size of

Earth); (2) the much greater distance from the sun of Mars than of Earth (hence less radiation received); and (3) the orbit of Mars is more eccentric, or elliptical, than that of Earth. These factors all have a fundamental bearing on a planet's atmosphere and weather system.

It has been generally established that a planet's size determines the strength of the internal heat sources, coming mainly from radioactive decay processes and gravitational energy released during accretion, that drive tectonic and volcanic activity. Even though, as a smaller planet. Mars sustained volcanic activity (as evident from *Viking* images), it was not imbued with the volcanic potential of Earth, which continues apace. Plate tectonics on Earth permits frequent access to internal heat sources, whereas *Viking* images indicate no recent evidence of plate tectonics; the entire crust appears to be a single plate. (Not full agreement among observers regarding this point.)

The greater distance of Mars from the sun obviously is a factor that must be built into any model of Mars. However, it may not be a major determining factor. With essentially general agreement that liquid water once existed on Mars and with suspicions that it may exist as ice in the regolith today (still very speculative), it follows that Mars received significant radiation from the sun in its early, formative periods, prior to losing its primitive atmosphere, to maintain water in the liquid state. (It has been theorized that once volcanic activity essentially abated on the planet, the atmosphere CO₂ level and hence greenhouse effect declined, causing cooling to the point where liquid water could no longer exist.) Kahn (Washington University) suggests that the surface pressure on the planet is so low today because CO₂ continued to be removed from the atmosphere and stored as carbonates by transitory pockets of liquid water. As interpreted by Haberle (NASA Ames Research Center), such pockets could have existed long after the global mean temperature had dropped below the freezing point; specifically, they could form as long as the surface pressure was sufficiently high to limit evaporation. Through the action of transitory water pockets the pressure was gradually reduced to its present value of 6.1 millibars. Haberle further observes that the small size of Mars probably had at least as much influence on its climate as has its distance from the sun. Moreover, the size of Mars has determined the fate not only of water and CO₂, but also of nitrogen, which is relatively scarce in the planet's atmosphere. The lower level of volcanic activity on Mars meant that less nitrogen was outgassed than on Earth. The smaller gravitational pull of Mars also made it easier for nitrogen to escape. (Although nitrogen does not have enough thermal energy to escape, it can acquire the necessary energy by dissociative recombination.)

Because of the eccentricity of Mars' orbit, the seasons on the planet are of unequal duration and intensity. (The Martian year is 687 Earth days long.) Perihelion (closest approach to the sun) occurs late in the southern hemisphere, making it 52 Earth days shorter than in the north. When at perihelion, Mars receives some 40% more solar radiation than when at aphelion. In terms of the Earth, this difference is only 3%. This asymmetry of seasons markedly affects the weather on Mars as it is known today, particularly influencing the cycles of CO₂, of water, and of dust (an important factor in the planet's current climate.) These cycle have been plotted by several researchers and are contained in the Haberle reference listed.

Obvious major differences in the weather of the two planets cast doubts on attempting to compare Earth models with those of Mars. The thin atmosphere of Mars today eliminates any consideration of a greenhouse effect. The absence of oceans eliminates the Earth's familiar hydrologic cycle. As pointed out later, these major differences in the way the martian atmosphere and "weather" system has evolved helps to explain some of the very unusual geologic features found in the Viking images.

Soil and Rocks

Nergal, Ares, and Mars were legendary names for a pinpoint of reddish light in the night sky, observed to move relative to the star field even in ancient times. Because of its color, Mars was an important part of the mythology of early civilizations, serving as an abode for gods of fire, war and terror in the minds of many populaces through the centuries. The ancients would not have been disappointed in the coloring (reddish-brown) of most of the Martian soil and rocks.

Steps taken in the operation of the surface sampler on the *Viking* landers is shown in Fig. 19. As pointed out by the X-ray analysis team,⁴ elemental analyses of fines in the martian regolith at the two widely separated landing sites were remarkably similar. At both sites, the uppermost regolith was found to contain abundant silicon and iron, with significant concentrations of magnesium, aluminum, sulfur, calcium, and titanium. The sulfur concentration is one to two orders of magnitude higher, and potassium ($<0.25\%$ by weight) is at least five times lower than the average for the Earth's crust. The trace elements strontium, yttrium, and possibly zirconium, were detected at concentrations near or below 100 parts per million. Pebble-sized fragments at VL-1 site were found to contain more sulfur than the bulk fines, and were thought to be pieces of a sulfate-cemented duricrust. It is interesting to note that no phosphorus was found on Mars.

Each *Viking* lander carried an energy-dispersive X-ray fluorescence spectrometer for elemental analysis of samples of the Martian surface. This composition is best interpreted as representing the weathering products of mafic igneous rocks. A mineralogic model, derived from computer mixing studies and laboratory analog preparations, has suggested that the martian fines could be an intimate mixture of about 80% iron-rich clay; about 10% magnesium sulfate (kieserite perhaps), about 5% carbonate (calcite?) and about 5% iron oxides (hematite, magnetite, maghemite, goethite?). The mafic nature of the fines, which appear to be distributed globally, and their probable source rocks seem to preclude large-scale planetary differentiation of an earthly nature. The samples were characterized by abundant red-colored fine material and scattered blocks of generally angular rocks. More diversity was found at VL-1 than at VL-2.

⁴Scientists with Martin-Marietta Aerospace Corp., NASA Langley Research Center, Pomona, College, the University of New Mexico, and the U.S. Geological Survey.

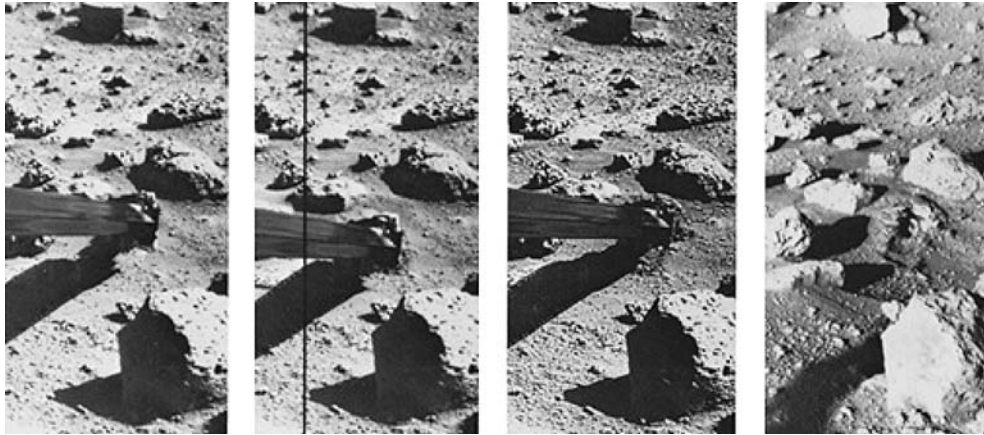


Figure 19. Operation of the surface sampler in obtaining martian soil was closely monitored by one of the Lander cameras because of the precision required in trenching a small area (8×10 inches; 20×25 centimeters) surrounded by rocks. The exposure of thin crust appeared in unique contrast with surrounding materials and became a prime target for organic analysis in spite of potential hazards. The large rock in the foreground is only 8 inches (20 centimeters) high. At left, the sampler scoop has touched the surface, missing the rock at upper left by a comfortable 6 inches (15 centimeters), and the backhoe had penetrated the surface about 0.5 inch (13 millimeters). The scoop was then pulled back to sample the desired point and (second view) the backhoe furrowed the surface, pulling a piece of thin crust toward the spacecraft. The initial touchdown and retraction sequence was used to avoid a collision between a rock (in the shadow of the arm) and a plate joining the arm and scoop. The rock was cleared by 2 to 3 inches (5 to 7.5 centimeters). The third picture was taken 8 minutes after the scoop touched the surface and shows that the collector head has acquired a quantity of soil. With the surface sampler withdrawn (right), the foot-long (0.3-meter) trench is seen between the rocks. The trench is 3 inches (7.5 centimeters) wide and about 1.5 to 2 inches (3.8 to 5 centimeters) deep. The scoop reached to within 3 inches (7.5 centimeters) of the rock at the far end of the trench. Penetration appears to have left a cavernous opening roofed by the crust and only about one inch (2.5 centimeters) of undisturbed crust separates the deformed surface and the rock. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

Terrain

The Martian terrain in the vicinity of the two *Viking* lander sites is well illustrated in latter part of this article. Of course, these sites represented only a portion of the planet. There are ten or more volcanoes prominent on Mars, with scientific estimates of their age ranging from 100 million years to 1 billion years. See Figs. 20 and 21.

Evidence of erosion, including dry channels resembling riverbeds and tributaries, has led many analysts to conclude that Mars may have had a warmer, more water-rich climate in the past. Photographic evidence from spacecraft indicates that the once-reported “canals” are mostly illusory and that the dark patchy markings once suspected to be vegetation along these canals, varying with the seasons, are in reality simply deposits of wind-blown dust that may be altered from time to time. That alterations can and do occur on Mars is shown by Fig. 22. With the benefit of a few years to assimilate observational data from

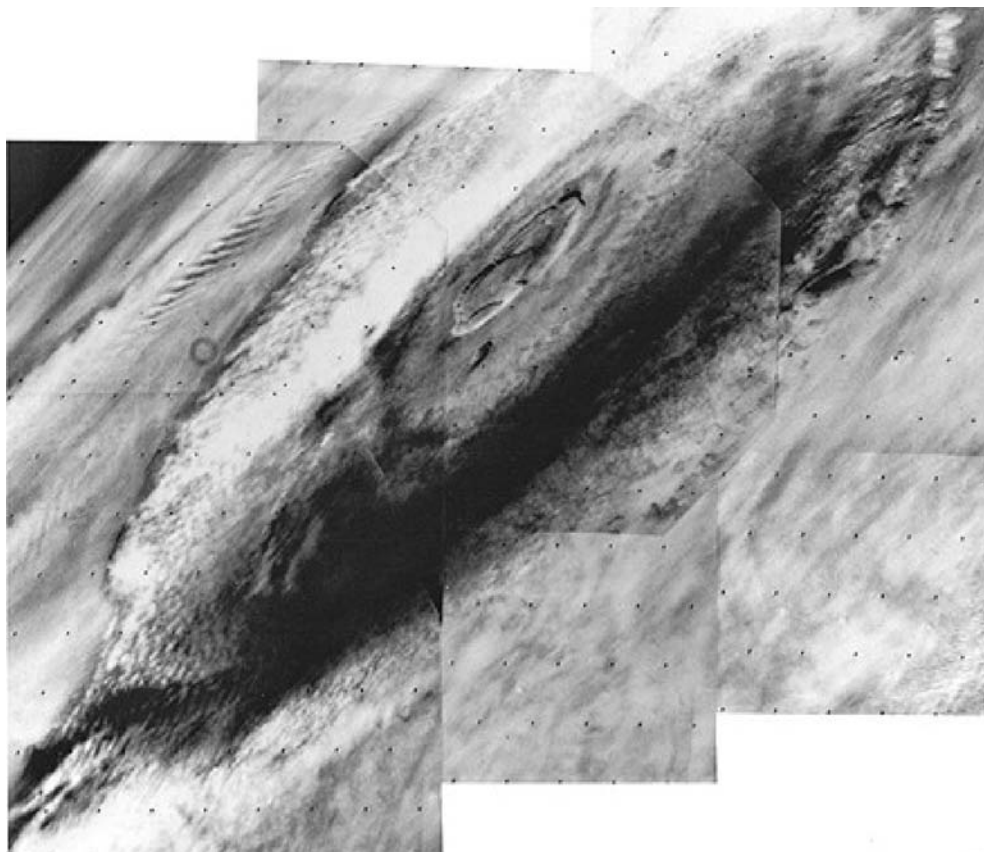


Figure 20. The great Martian volcano Olympus Mons was photographed by the *Viking 1 Orbiter* on July 31, 1976 from a distance of 8000 kilometers (5000 miles). The 24-kilometer-high (15-miles) mountain is seen in mid-morning, wreathed in clouds that extend up the flanks to an altitude of about 19 kilometers (12 miles). The multi-ringed caldera (volcanic crater), some 80 kilometers (50 miles) across, pushes up into the stratosphere and appears cloud-free at this time. The cloud cover is most intense on the far western side of the mountain. A well-defined wave cloud train extends several hundred miles beyond the mountain (upper left). The planet's limb can be seen at the upper left-hand corner of the view. It also shows extensive stratified hazes. The clouds are thought to be composed mainly of water ice condensed from the atmosphere as it cools while moving up the slopes of the volcano. In the martian afternoon, the clouds develop sufficiently to be seen from earth and it is known that they are a seasonal phenomenon largely limited to spring and summer in the northern hemisphere of the planet. Olympus Mons is about 600 kilometers (375 miles) across at the base and would extend from San Francisco to Los Angeles. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

Mariner and the *Viking* missions, the findings of Pieri et al. (1980) are exceptionally interesting. As explained by Pieri, *Mariner 9* orbital reconnaissance mission discovered ubiquitous valley networks in heavily cratered terrain. Branching valley networks throughout the heavily cratered terrain of Mars exhibit no compelling evidence for formation by rainfall-fed erosion (one of the



Figure 21. Fine detail in the interior of a martian crater can be seen in this photo taken by *Viking 1* of an area near the *Viking 2* landing site. The crater (on the left margin of the view) is about 40 kilometers (25 miles) in diameter and shows many features found in lunar craters. The central portion is crossed by numerous cracks. Similar features are seen at the huge lunar impact basin, Orientale. Their origin is unknown, but it has been suggested that the cracks were formed either by consolidation of lava that filled the crater after it formed, or by fallback from the impact process. Alternatively, the cracks may have formed long after the impact event by uplift of the crater floor. Between the cracked terrain and the crater rim is a region of chaotic debris. Beyond the rim there is no evidence of an ejecta blanket (rock material which is blasted from the crater by the shock of the impacting meteorite). The ejecta blanket is presumably overcovered by later deposits. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

popular hypotheses). Rather, the networks are diffuse and inefficient, with irregular tributary junction angles and large, undissected intervalley areas. The deeply entrenched canyons, with blunt amphitheater terminations, cliff-bench wall topography, lack of evidence of interior erosion of flow, and clear structural



Figure 22. Changes observed by the *Viking Landers* over a period of time included water-ice snow seen by VL-2 during the winter at Utopia Planitia, and a thin dust layer deposited at both sites during the dust storms of 1977. As shown here, a change occurred by Chryse Planitia over a 4-day period in September 1978. Top photo is the “before” and bottom photo is the “after” view. Change (A) appears as a small circle-like formation on the side of a drift in the lee, or downward, side of “Whale Rock.” This is believed to have been a small-scale landslide of an unstable dust layer which had accumulated behind the rock. Interpretation of this feature would be difficult without an earlier change (B) near “Big Joe,” a slump. The new slump is observed approximately 25–35 meters (82–115 feet) from the lander craft and just under 1 meter (3.3 feet) across. This slumping was probably initiated by the daily heating and cooling of the surface by solar radiation. More importantly, it is now believed that, based upon the repeated occurrence of such slumping features, a dust layer which overlies the surface may, in fact, be redistributed fairly regularly during periods of high wind activity. (NASA; *Jet Propulsion Laboratory, Pasadena, California.*)

control, suggest headward extension by basal sapping.⁵ The size-frequency distributions of impact craters in these valleys and in the heavily cratered terrain that surrounds them are statistically indistinguishable, suggesting that valley formation has not occurred on Mars for billions of years.

Pieri observes that the branching and coalescent character of the channels provoked immediate comparison to terrestrial riverine networks produced by fluvial⁶ erosion, a process driven primarily by rainfall. Liquid water, however, cannot now exist at the surface of Mars for more than a few minutes owing to a very low atmospheric pressure and very cold temperatures during most of the year at most latitudes. It has been suggested that these features, if formed by processes similar to those that operate in the formation of earthly river systems, are relics of a more clement epoch. Thus, a major problem in the study of Mars is whether the valley networks could have evolved under current surface conditions, or whether a major shift in Martian climate occurred.

There are certain unifying characteristics of Martian valleys. (Valleys are distinguished from channels by the absence in the former of direct evidence of fluid erosion often found in the latter.) There is no clear evidence (streamlined obstacles, interior channels) of direct fluid erosion in any Martian valley. It is possible that such features are too small to be observed by present instrumentation, although *Viking Orbiter* images as small as 100 meters (328 feet) can be resolved. Walls of the valleys are typically rugged and clifflike, with some debris accumulation and talus, and the floors are generally flat. Mantling by materials of eolian and volcanic origin is common. Some valleys display cliff-bench interior topography, similar in character and scale to features in the Grand Canyon of the Colorado River. The most striking morphological characteristics, however, are the presence of steep-walled, cusped terminations at the heads of the smallest tributary valleys. These steep-walled, amphitheater terminations suggest headward extension (sapping) by basal undermining and wall collapse, as in the predominant mode of headward extension for many earthly canyons. A variety of Martian terrain is shown in Figs. 23–26.

Martian valley networks lack the dendritic pattern so common to terrestrial streams. The Martian valley patterns show remarkable parallelism and lack of tributary competition for undissected intervalley terrain, and thus appear diffuse compared to terrestrial systems. Viewed from spacecraft, terrestrial drainage systems have a fine, filigreed texture, whereas Martian systems appear coarse.

Pieri has suggested that the valleys were formed on Mars during an ancient epoch by erosional processes involving not rainfall, but the movement of groundwater and its participation as a liquid or a solid in the undermining of less competent strata, causing progressive headward collapse. These processes, combined with modification by impact and eolian (wind) processes, have produced the degraded valleys seen on Mars today.

Even a brief description of the physical features of Mars is not complete without mention of the so-called “Spokane Flood” concept. As summarized by

⁵The undercutting, or breaking away of rock fragments, along the headwall of a cirque (semicircular, amphitheater-like, or armchair-shaped hollow of nonglacial origin), due to frost action at the bottom of the bergschrund (a deep and often wide gap or crevasse).

⁶Of or pertaining to a river or rivers—produced by the action of streams.



Figure 23. The sinuous rille (a relatively long, narrow, trench- or cracklike valley) at the top of this mosaic of 8 photos is believed by some scientists as indicative of flooding of the high plateau in the vicinity of an alternative landing site (known as Capri) for *Viking Lander 2*. In the foreground is a valley that may have been caused by downfaulting of the martian crust. The hummocks (rounded or conical knolls or mounds of comparatively small elevation) on the valley floor look like chaotic terrain. Some scientists have suggested that the subsidence may be partially caused by melting of the subsurface ice. The large areas of the collapsed terrain show the regional extent of this phenomenon. These views were taken by *Viking 1* on July 3, 1976 from a range of 2300 kilometers (1400 miles) and cover an area of about 300×300 kilometers (180×180 miles). South is toward the top as seen from the spacecraft. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

Baker (1978), in a series of papers published between 1923 and 1932, J.H. Bretz described an enormous plexus of proglacial stream channels that eroded into the loess and basalt of the Columbia Plateau in eastern Washington state. Bretz argued that this region (which he termed the Channeled Scabland) was the product of a cataclysmic flood, which he called the Spokane flood. Considering the nature and vehemence of the opposition to his hypothesis, which at one time was considered highly imaginative, its eventual scientific verification constitutes a fascinating episode in the history of modern science. The discovery of possible catastrophic flood channels on Mars has given new relevance to Bretz's insights. The connection between Bretz's proposal and parts of the Martian surface is well developed by Baker.

Volcanism. As readily apparent from *Viking* images, volcanism on Mars was widespread. According to Lucchitta (U.S. Geological Survey), volcanism has formed enormous shields, large composite cones, lobate lava flows, and possibly small cones and pseudocraters. Flood basalts similar to those filling lunar maria may have resurfaced ridged highland plateaus. Large deposits of pyroclastic



Figure 24. View taken by *Viking 1* on July 3, 1976 from a range of 2000 kilometers (1240 miles), looking southward across Valles Marineris. This huge equatorial canyon is about 2 kilometers (1.2 mile) deep. The area shown is 70 kilometers (43 miles) by 150 kilometers (94 miles). Aprons of debris on the canyon floor indicate how the canyon may have enlarged itself. The walls appear to collapse at intervals to form huge landslides that flow down and across the canyon floor. Linear striations on the landslide surface show the direction of flow. On the canyon's far wall in this view, one landslide appears to have ridden over a previous one. Streaks on the canyon floor, aligned parallel to the length of the canyon, probably are evidence of wind action. Layers in the canyon wall indicate that the walls are made up of alternate layers of lava and ash or wind-blown deposits. (NASA; Jet Propulsion Laboratory, Pasadena, California.)



Figure 25. A view obtained by *Viking 1* on July 8, 1976 showing what appear to be fault zones in the martian crust in an area two degrees south of the equator. The fault valleys are widened by mass wasting and collapse. Mass wasting is the downslope movement of rocks due to gravity (possibly hastened by seismic shaking if present). (NASA; Jet Propulsion Laboratory, Pasadena, California.)

material may also exist, although their presence is controversial. Dark patches are common on the planet as they are on Earth's moon, where they have been interpreted as pyroclastic materials. The association of dark patches with pyroclastic volcanism on Mars has been largely overlooked, because most dark



Figure 26. Mosaic of martian surface made by *Viking Orbiter 1* over a period between August 4 and 9, 1976. The area is centered at 17°N, 55°W, to the west of the Viking 1 Lander site in Chryse Planitia. Just to the west of this area are the plains of Lunae Planum. The terrain shown in this view slopes from west to east with a drop of about 3 kilometers (1.9 mile). The channels are a continuation of those to the west of the VL-1 landing site and, to some scientists, suggest a massive flood of waters from Lunae Planum, across this intervening cratered terrain, and into the general region of the VL-1 landing site. In several cases, it will be noted that channels cut through craters; in others, the craters are clearly later than the assumed flood and are superimposed in the channels. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

patches are inside craters and were obviously accumulated by wind; the possibility was neglected that some Martian dark patches, like lunar ones, may reflect pyroclastic vents. Lucchitta describes dark patches in Valles Marineris that may be such vents and may reflect young mafic volcanism. The evidence for past volcanism on Mars is commonly accepted, but none has been documented in the Valles Marineris equatorial rift system. A recent survey of the troughs in this valley revealed dark patches that are interpreted to be volcanic vents. The configuration and association of these patches with tectonic structures suggest that they are of internal origin; their albedo and color ratios indicate a mafic composition; and their stratigraphic position, crispness of morphologic detail, and low albedo imply that they are young, perhaps even recent. If this volcanism is indeed as young as it seems, Mars has been an active planet throughout most of its history.

Case for an Early “Wet” Mars. Certain features of the Martian surface, as observed by the Viking missions, have continued to intrigue and confuse analysts of the data returned from Mars. Included are ancient valleys, channels, and what appear from images to be tributary systems. What appear to be numerous extinct volcanoes and meteoritic craters over which more recent geologic features have been superposed are found in the images. For many of these features, the presence of water in relatively large quantities on the planet during its earlier phases offers the most tempting solution. Many scenarios have been developed. For example, some scientists at the Jet Propulsion Laboratory (Pasadena, California) suggested, at a symposium of the Lunar and Planetary Institute (see reference listed), that lakes, or a shallow sea or ocean, may have encompassed as much as 10 to 15% of the Martian surface and of a generous portion of the northern plains from 2 to 3 billion years ago. Timothy Parker (JPL) estimates that the water would have been about 100 meters deep (or less), making the sea’s volume equivalent to a layer of about 10 meters deep covering the globe. This volume of water in a hypothetical sea would equal all the water that some geochemists have allowed for the entire planet. But, all do not agree. Michael Carr (U.S. Geological Survey) reported that the latest estimates of the amount of water hidden beneath the surface may be several times greater. Some investigators believe that the surface of Mars, something comparable to Earth’s moon, is made up of rubble and porous soil well capable of storing ice, water, or brine. Stephen Clifford (Lunar and Planetary Science Institute) estimates that this megaregolith has the capacity to hold water equivalent to a global layer 200 to 500 meters deep. The main remaining question, of course, is how much of that capacity is actually filled? Certainly, the unanswered question of the amount of water that was and still may be trapped on the planet is central to preparing a satisfactory model of the planet.

Age Determination. Mars has been mapped extensively by *Mariner 9* and later by the *Viking* missions. One major goal in planetary science is to determine the chronology of development of the surfaces of the terrestrial planets, particularly Mars. As indicated in an excellent paper by Neukum and Wise (1976), cratering links to lunar time suggest that Mars died long ago. Fortunately, for the purpose of age determination from photographs, Mars is impact-cratered. Differences in impact crater frequencies at different sites reflect differences in age. Two attempts have been made to determine absolute age for Mars from its

measured crater frequencies, based on extrapolations from the cratering chronology of the lunar surface (Hartman, 1973; Soderblom et al., 1974). Unfortunately, a straightforward comparison of martian and lunar crater frequencies does not necessarily yield true ages; relative impact rates and the time dependence of the martian cratering rate are not known; and it is not certain whether the same meteoroid population bombarded both planets.

At *Mariner 9* resolution, the impact crater production size-frequency distribution of Mars is generally similar to that of the moon for crater diameters in the range 0.8 to 50 kilometers (0.5 to 31 miles), and it appears to have been relatively stable through time. The lunar and Martian crater curves can be brought into near coincidence by a diameter shift appropriate to reasonable impact velocity differences between bodies hitting Mars and the Moon. This indicates that a common population of bodies impacted both planets and suggests the same or a very similar time dependence of impact flux. Constraints on relative lunar and Martian fluxes can be obtained by comparing crater frequency data for the lunar and martian highlands and for Mars' satellite Phobos.

These cratering constraints, as pointed out by Neukum and Wise, provide the basis for a tentative Martian time scale derived from lunar data. Previous time scales have painted a picture of a disorderly planetary evolution of Mars, punctuated by a strange pulse of Tharsis Ridge tectonic and volcanic activity late in geological history. The new scale suggests a much more orderly evolution with Mars, like the Moon, winding down most of its major planetary tectonic and volcanic disturbances in the first 1.5 billion years of its history. By 2.5 billion years ago the volcanic-tectonic era on Mars had ended.

Other Physical Characteristics of Mars

Doppler radio-tracking data have provided detailed measurements for a Martian gravity map extending from 30°S to 65°N latitude and through 360° of longitude. The feature resolution is approximately 500 kilometers (310 miles), revealing a huge anomaly associated with Olympus Mons, a mascon in Insidis Panitia, and other anomalies correlated with volcanic structures. Olympus Mons has been modeled as a disk of 600-kilometer (372-mile) surface area, having a mass of 9.7×10^{21} grams. The very large Olympus Mons anomaly should have a very significant impact on geophysical modeling of the planet. Similarly, the Elysium anomaly and the Insidis mascon should place constraints on the internal structure. Gravity in the southern hemisphere remains poorly resolved.

A three-axis short-period seismometer was delivered to the surface of Mars by *Viking Lander 2* on September 3, 1976. Noise background correlated well with wind gusts. Data returned to earth indicated that Mars is a very quiet body.

The amounts of magnetic particles held on the reference test chart and backhoe magnets on *Viking Landers 1* and 2 were comparable, indicating the presence of an estimated 3–7% (weight) of relatively pure, strongly magnetic particles in the soil. It is argued that the results indicate the presence, now or originally, of magnetite, which may be titaniferous.

Dust Devils on Mars. Several scientists, after studying Viking data, have reported the existence of dust devils (columnar, cone-shaped, and funnel-shaped

clouds rising 1 to 6 km above the surface) on Mars. Dust devils result from atmospheric conditions that occur close to the ground and are, therefore, sensitive to surface topography. Dust devils on Mars may be responsible for the initiation of large dust storms on the planet and for increasing the general atmospheric dust content.

Dust devils, as observed by Thomas and Gierasch (Cornell University), have meteorological as well as geological significance. Fluid motions in an atmospheric boundary layer can be driven either by stresses due to the mean wind (forced convection) or by buoyancy due to heating of the gas adjacent to the surface (free convection). Dust devils are an example of the latter. On Earth, large-scale eolian transport is generally due to forced convection. The investigators report that moderate to high winds characterize forced convection, and on Mars, where the atmospheric density is only about 1% of that on Earth, it is estimated that winds must exceed about 25 to 40 meters sec^{-1} to initiate soil movement.

One of the major geologic processes on Mars is the entrainment and transportation of dust by winds. Observations on the genesis and development of local and global dust storms on Mars are sparse.

Tornadolike Tracks on Mars. Some images from the *Viking Orbiter* reveal well defined, dark filamentary lineations in numerous locations on the Martian surface. On Earth, tornadic-intensity vortices commonly leave distinctive tracks whose appearance is similar to that of the Martian lineations. A high-resolution imaging system, as proposed for the *Mars Observer* mission, could resolve these ground tracks. The filamentary lineations, as reported by Grant and Schultz (Brown University) are from 2 km to at least 75 km long and less than 1 km wide. Most are straight to curvilinear, and some have obvious nontopographically initiated gaps in their path. The visible occurrence of the lineations appears to be seasonal. In the southern hemisphere, they were visible (from *Mariner 9*) only from midsummer into early fall. After formation, they were rapidly modified and were no longer visible by midfall. In the northern hemisphere, lineations appear from early to midsummer. By late summer, these lineations also become smeared and faint.

Natural Laser Phenomenon Noted on Mars. Based upon observations made with the Goddard infrared heterodyne spectrometer during the period of January to April 1980, when the planet was near opposition, astronomers M.J. Mumma and colleagues (NASA-Goddard Space Flight Center) and D. Zipoy (University of Maryland) noted natural gain amplification in the mesosphere of Mars, probably representing the first definite identification of a natural infrared laser. Natural microwave amplifiers (masers) are abundant in interstellar clouds and some circumstellar shells, primarily among the rotational level populations of certain molecules, such as OH, SiO, and H₂O, but no optical lasers in nature had previously been observed. As reported by Mumma et al., many examples of natural nonthermal optical emission have been found, such as the infrared and ultraviolet auroras or the day glows of Earth, Jupiter, Mars, and Venus. Details are reported in the reference listed.

Pole Wandering and Crustal Shifts on Mars

Careful study of *Viking* images has revealed a number of features of the planet that are very difficult to explain. For example, regions at the planet's equator

seem once to have been near a pole. As observed by Schultz (see reference), in certain areas of the surface, erosion appears to have occurred at a very low rate (perhaps less than a millimeter in a million years). But, in other areas, at the same latitude, there are regions that have been heavily stripped and etched by the wind. Also, very old networks of narrow valleys, once cut into the surface by flowing water, suggest a warm climate, although such networks are seen within 10 degrees of the southern polar ice cap. While many details remain to be worked out, Schultz suggests that one hypothesis may explain all or most of the contradictions: the orientation in space of the Martian crust has not always remained the same throughout geologic time, but rather, it has shifted with respect to the planet's axis of spin. This would require that the spin axis, which intersects the planet's surface at the north and south poles, would appear to have wandered over the planet's crust. This would indicate that certain regions of the crust, presently far from the poles, may have been at some time in the past within the polar regions. In introducing a detailed paper on this topic, Schultz observes that if Mars had undergone polar wandering, then martian geology may have to be viewed in the context both of a dynamically changing planet like the Earth and of a stable, rigid body like the Earth's moon. In this sense, the Martian equivalent of plate tectonics might simply be the movement of the entire lithosphere, the solid outer portion of the planet, as one plate.

Martian Satellites

Mars has two satellites, Phobos, the inner and larger of the two moons, and Deimos. These satellites were visited during the *Viking* missions.

Phobos. This satellite revolves around Mars in an orbit of about 9330 kilometers (5800 miles) from the center of Mars (some 5950 kilometers; 3700 miles above the planet's surface). The diameter was stated in the introductory section of this entry. Its orbital period is 7 hours, 40 minutes. Because its orbital period is in the same direction as, but is less than, that of Mars, it rises in the west and sets in the east as seen from Mars. Phobos is heavily cratered and dark in color, of a material resembling carbonaceous chondrite meteorites. A system of grooves, possibly marking fractures, is associated with the largest crater, Stickner, which is about 10 kilometers (6.2 miles) across.

Viking Orbiter 1 flew within 480 kilometers (300 miles) of Phobos to obtain the view given in Fig. 27. A view much closer to the satellite is given in Fig. 28. A considerably later view, made in 1978, is given in Fig. 29.

Deimos. This satellite revolves around Mars in an orbit about 23,000 kilometers (14,260 miles) from the center of Mars. Five close flybys, within 1000 kilometers (620 miles) of Deimos, were made in October 1977. The closest encounter was on October 5, 1977 when the spacecraft passed within 50 kilometers (30 miles) of the moon's surface. Images indicated that the surface of Deimos differs considerably from that of Phobos. Deimos has many craters, but appears smoother than Phobos. See Fig. 30. With reference to the peculiar blocks observed on Deimos, which are visible in the illustration, Duxbury and Veverka (1978) suggest: "If the bright patches and blocks represent ejecta, then it is puzzling why apparently so much of it was retained by such a small satellite and

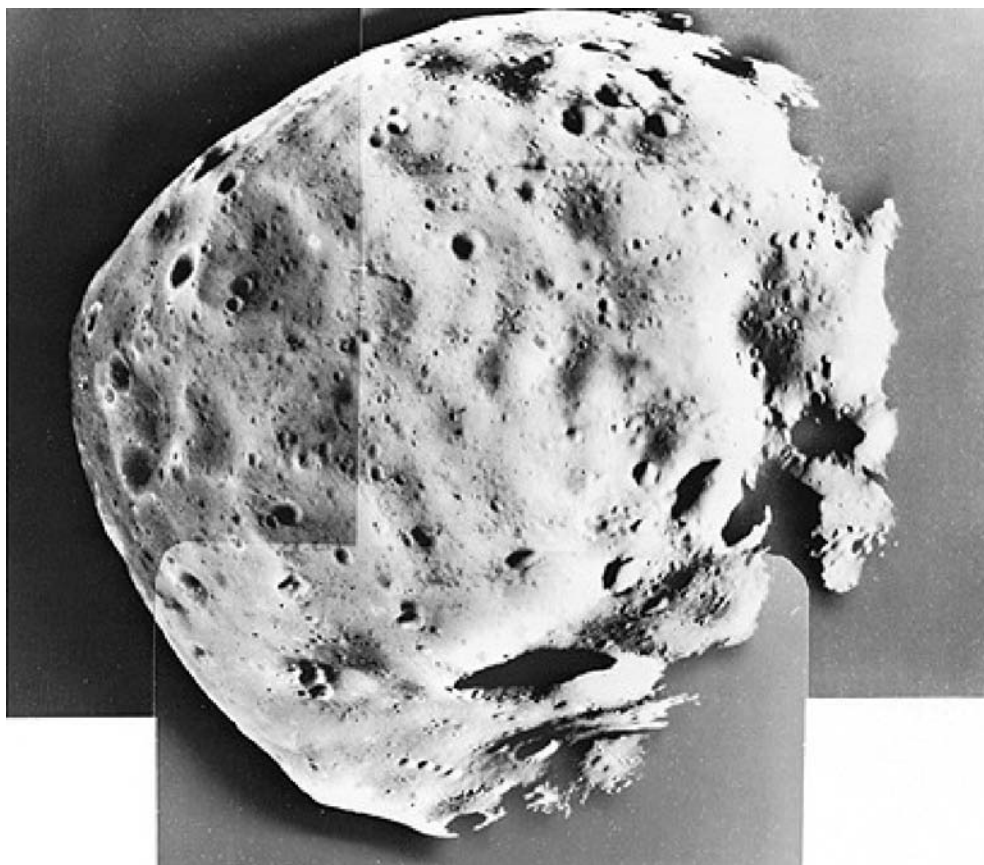


Figure 27. View of Phobos taken by *Viking Orbiter 1* from a distance of 480 kilometers (300 miles). This mosaic of 3 pictures was made in February 1977. As seen here, Phobos is about 75% illuminated and is about 21 kilometers (13 miles) across and 19 kilometers (11.8 miles) from top to bottom. North is at top. The south pole is within the large crater (Hall) with a diameter of 5 kilometers (3.1 miles) and will be noted at bottom center where the pictures overlap. Some features as small 20 meters (65 feet) across can be seen. Remarkable features include striations, crater chains, a linear ridge, and small positive features which appear to be resting on the surface. A long linear ridge is seen starting near the south pole and extending to the upper right. A very sharp wall at the intersection of two craters (about 1 kilometer; 0.6 mile across) is seen along this ridge at right. A series of craters runs horizontally in the picture which is parallel to the orbit plane of Phobos. These crater chains are commonly associated with secondary cratering by ejecta from larger impacts. A surprising discovery has been made of what apparently resembles hummocks or small positive features. These features, primarily seen near the terminator (right), are about 50 meters (165 feet) in size and may be surface debris from previous impacts. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

why the process seems to be so much more efficient on Deimos than on Phobos. It is conceivable that the very close proximity of Phobos to Mars makes it easier for impact ejecta to escape from the inner satellite, but the mechanics of such a preferential process remain to be worked out.”



Figure 28. *Viking Orbiter 1* took this close-up picture of Phobos from a range of 120 kilometers (75 miles) on February 20, 1977. This is the closest range at which a spacecraft has photographed the tiny satellite. At that range, Phobos is too large to be captured in a single frame. This picture covers an area 3×3.5 kilometers (1.86×2.17 miles). A single picture element is about 3 meters (7.5 feet) across. However, the high relative speed of Orbiter 1 and Phobos caused some image smear so that the smallest surface feature identifiable is between 10 and 15 meters (32 and 49 feet). The picture shows a region in the northern hemisphere of Phobos that has striations and is heavily cratered. The striations, which appear to be grooves rather than crater chains, are about 100 to 200 meters (328 to 656 feet) wide and tens of kilometers long. Craters range in size from 10 meters (32 feet) to 1.2 kilometers (0.75 mile) in diameter. The surface of Phobos appears similar to the high-land regions of the earth's moon, which also is heavily cratered and an ancient terrain. The dark region above the limb of Phobos is an artifact of processing and does not indicate an atmosphere. (NASA; Jet Propulsion Laboratory, Pasadena, California.)



Figure 29. This view of Phobos was made by *Viking Orbiter 1* on October 19, 1978 at a range of 612 kilometers (379 miles) during the spacecraft's 854th revolution of Mars. This view was made just before Phobos entered the shadow of the planet. The photomosaic shows the front side of Phobos which always faces Mars—from about 10° below the equator with north at the top. Stickney, the largest crater on Phobos (10 kilometers; 6.2 miles across) is at the left near the morning terminator. Linear grooves coming from and passing through Stickney appear to be fractures in the surface caused by the impact which formed the crater. Two earlier new encounters with Phobos brought *Viking Orbiter 1* within close range of the satellite, but had not provided scientists with good opportunities to observe Stickney as well. This view provides new high-resolution coverage of the front side of Phobos (approximately 19×22 kilometers; 11.8×13.6 miles as seen here) as well as the highest resolution yet achieved of the western wall of Stickney. Kepler Ridge is casting a shadow in the southern hemisphere which partially covers the large crater (Hall) at the bottom. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

Additional Post-Viking Mission Studies and Hypotheses

Further studies of the Viking information and observations made from Earth in recent years have posed interesting new questions pertaining to Mars.

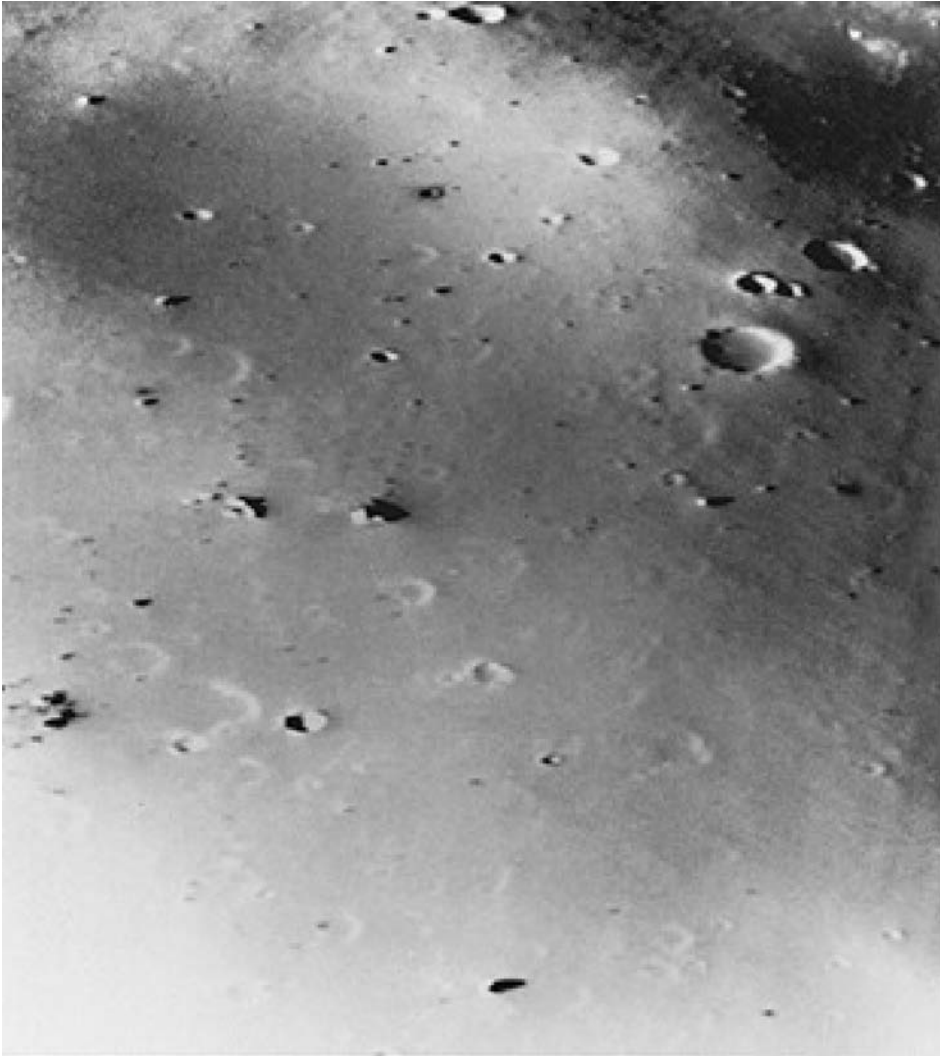


Figure 30. View of the Martian moon Deimos taken on October 15, 1977 when *Viking Orbiter 2* passed within 50 kilometers (30 miles) of the satellite. The picture covers an area 1.2×1.5 kilometers (0.74×0.93 mile) and shows features as small as 3 meters (10 feet). Deimos is saturated with craters, but a layer of dust appears to cover craters smaller than 50 meters (165 feet) in diameter, making Deimos look smoother than the other Martian moon, Phobos. Boulders as large as houses (10 to 30 meters; 33 to 100 feet) across are strewn about the face. It is suggested that these objects may be blocks ejected from nearby craters. The spacecraft would have been clearly visible to an observer standing on the surface of Deimos. (NASA: Jet Propulsion Laboratory, Pasadena, California.)

Radar Images of Mars. In late 1991, D.O. Muhleman and a team of researchers (California Institute of Technology) conducted aperture synthesis mapping of Mars by using the Very Large Array (VLA) in New Mexico as the imaging instrument to detect continuous wave signals transmitted at 9.5 GHz (3.5 cm) from the Jet Propulsion Laboratory (JPL) 70-meter antenna in

Goldstone, California. (See also Deep Space Network, Evolution of Technology.) Summary of the project: "The surface of Mars was illuminated with continuous wave radiation. The reflected energy was mapped in individual 12-minute snapshots with the VLA in its largest configuration; fringe spacings as small as 67 km were obtained. The images reveal near-surface features, including a region in the Tharsis volcano area, over 2000 km in east-west extent, that displayed no echo to the very low level of the radar system noise. This feature (called *Stealth*) is interpreted as a deposit of dust or ash with a density less than about 0.5 grams/cubic centimeter and free rocks larger than 1 cm across. The deposit is envisioned to be several meters thick and may be much deeper. The strongest reflecting geological feature was the south polar ice cap, which was reduced in size to the residual south polar ice cap at the season of observation. The cap image is interpreted as arising from nearly pure carbon dioxide or water ice with a small amount of Martian dust (less than 2 percent by volume) and a depth greater than 2 to 5 meters. Only one anomalous reflecting feature was identified outside of the Tharsis region, although the Elysium region was poorly sampled in this experiment and the north pole was not visible from Earth." More detail is given in reference listed.

Radar Detection of Phobos. During the exceptionally close approach of Mars to Earth in the autumn of 1988, the Goldstone 70-meter antenna was used as a radar telescope to observe Phobos. A total of 117 transmit/receive cycles were completed. Radar echoes from the Martian satellite provided information about the object's surface properties at scales near the 3.5-cm observing wavelength. In summary, "Phobos's surface apparently resembles those of many (if not most) large, C-class asteroids in terms of bulk density, small-scale roughness, and large-scale topographic character, but differs from the surfaces of the moon and at least some small, Earth-approaching objects. Additional 3.5-cm and 13-cm radar observations of asteroids, comets, and the martian satellites can clarify these relations."

Simulating the Surface of Phobos. Researchers (University of Arizona, Lunar and Planetary Laboratory) in recent years have been attempting to simulate how certain distinct features of Phobos may have been formed and as imaged by the Viking orbiters. Lines (rows) of comparatively small craters resemble a beaded chain unlike other features found in the planetary system thus far. An initial hypothesis described the features as being formed out of secondary ejecta—that is, debris resulting in a crater-causing impact. In attempting to duplicate the unusual feature in the laboratory, the researchers have developed an apparatus consisting of a pair of narrow, rigid glass plates and a variety of materials, including expanded vermiculite, silica sand, and small glass spheres. Intense interest in unusual surface conditions is not new in connection with planetary space explorations. Ponder, for example, the variety of scientific opinions that were expressed prior to touchdown of *Surveyor*, the first unmanned spacecraft to land on the Moon. From their work to date, the researchers have suggested that Phobos may be covered with a regolith some 300 meters thick!

A Peopled Mission to Mars. Although, in the early 1990s it was delayed for want of funding and other political considerations (due in part to some lack of interest on the part of the public) an *Apollo*-type mission to Mars is in the advanced planning phase. Conservative scientists have suggested, however, that

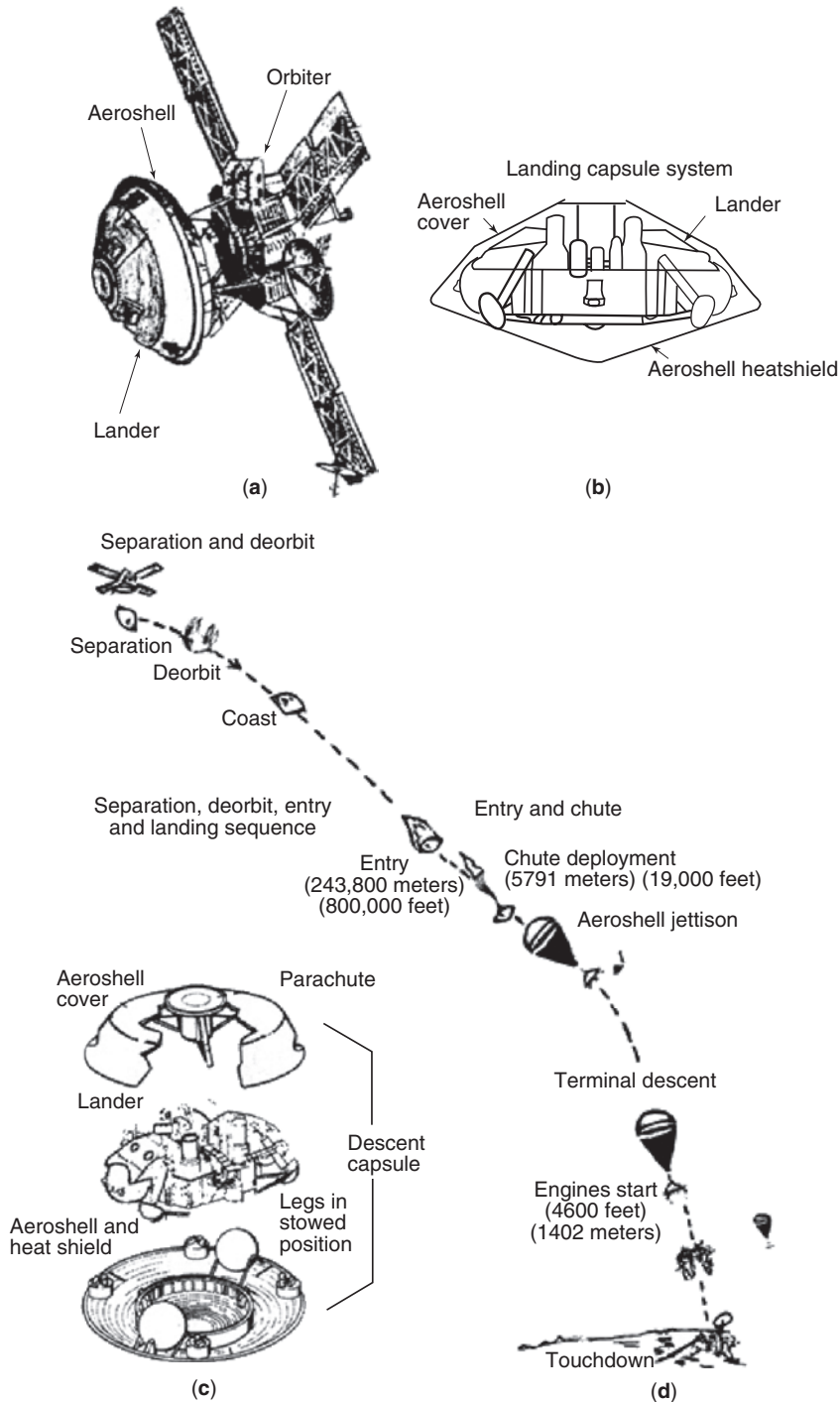


Figure 31. Principal subsystems of the *Viking* spacecraft: **(a)** Orbiter and Lander linked together as they travel through space; **(b)** landing capsule system; **(c)** aeroshell cover, parachute, and descent capsule; and **(d)** separation, deorbit, entry, and landing sequence. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

another *Viking*-type venture may be the safest and most sensible step prior to putting human lives at risk. Probably the most intense interest in a fully blown venture stemmed from Russian scientists, who have been preparing for a “Soviet” Phobos mission. A major concern has been and continues to be that of prolonged crew interest and mental and physical reactions to a sojourn in space that would require a currently calculated minimum of 15 months from launch to return on Earth. This has been referred to as the “Sprint” mission. For example, if it were assumed that the mission would leave Earth on 19 November 2004, it would reach at Mars on 30 July 2005 and depart Mars on 20 August 2005, returning to Earth on 2 February 2006. Time of stay on Mars would be less than 1 month and fuel costs would be at a maximum. A much longer (31 months) mission could be much more fuel efficient, and the stay on Mars could be considerably longer.

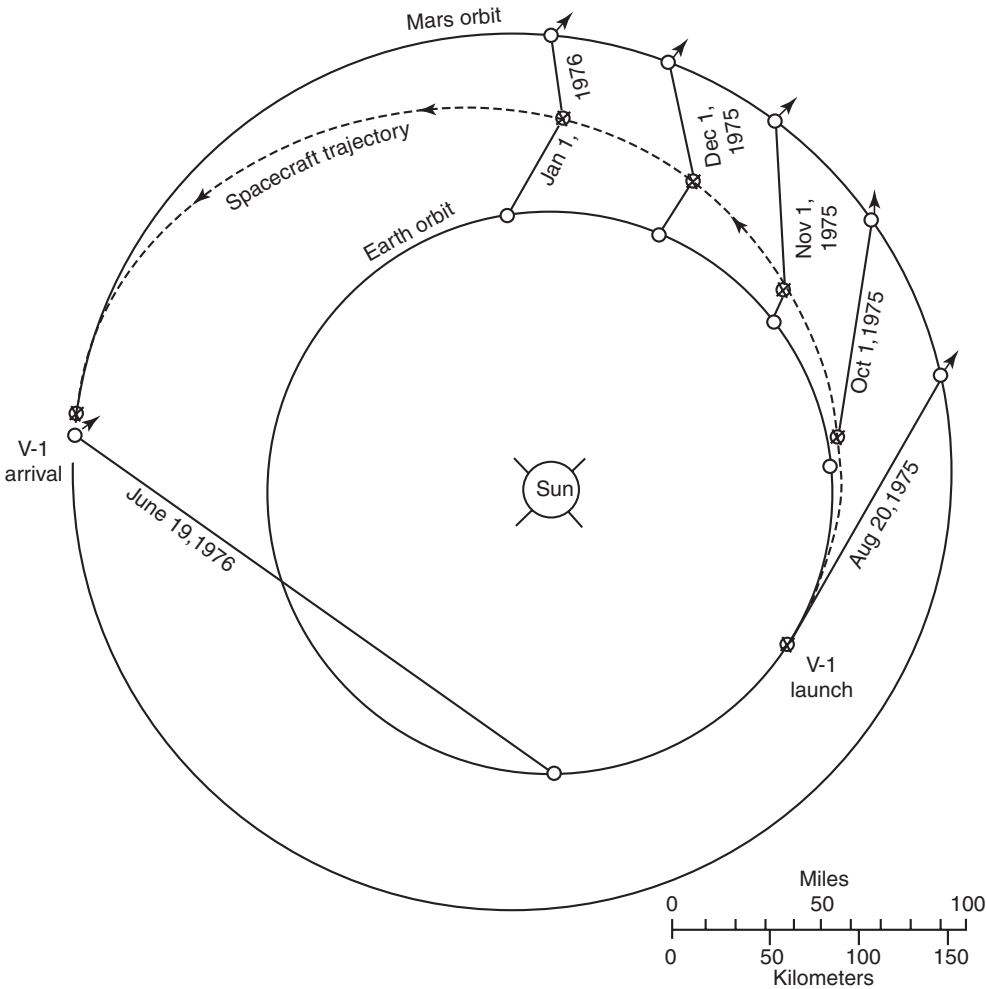


Figure 32. Trajectory followed by *Viking 1*. Dates given show relationship of earth, spacecraft, Mars, and the sun at specific times.

The Viking Missions to Mars

Two identical spacecraft, *Viking 1* and *Viking 2*, were launched in 1975 to explore Mars. In actuality, there were four spacecraft in all—a *Viking 1* orbiter and lander and a *Viking 2* orbiter and lander. Each orbiter and lander traveled together as one unit to rendezvous with Mars. The principal subsystems of the Viking spacecraft and separation, deorbit, entry, and landing sequences are shown in Fig. 31. The principal *Viking* events occurred as follows:

	<i>Viking 1</i>	<i>Viking 2</i>
Date of launch	August 20, 1975	September 10, 1975
Placed in elliptical orbit around Mars	June 6, 1976	August 7, 1976
Touchdown of Lander	July 20, 1976	September 3, 1976

Viking 1 traveled nearly 676 million kilometers (420 million miles) and *Viking 2* nearly 713 million kilometers (443 million miles) in their heliocentric Mars

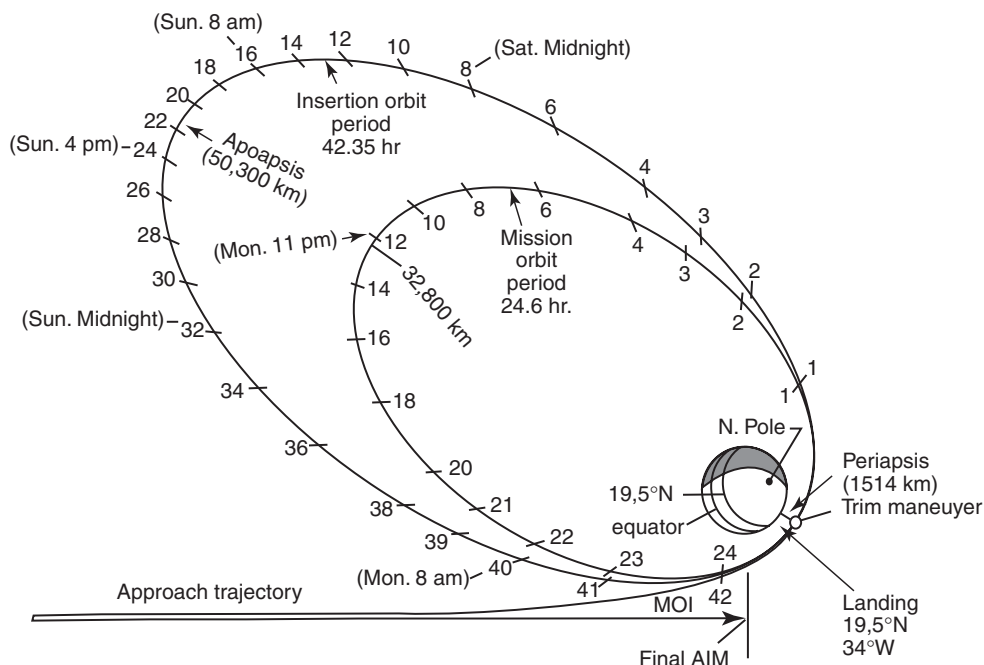


Figure 33. Orbit geometry for the insertion and mission orbits. *Viking 1* completed only one revolution on the insertion orbit before the trim maneuver placed it on the mission orbit. The tick marks indicate spacecraft flight hours with periapsis as the zero point. (Periapsis = the orbital point nearest the focus of attraction; apoapsis = the farthest point.) Additional information at selected points along the insertion orbit indicate where the spacecraft was that day relative to Earth Pacific Daylight Time. A complete revolution of Mars on the mission orbit required 24.6 hours, the length of a martian day. The orbit was synchronized with the landing site in that the spacecraft passed over the site once each day near periapsis, allowing maximum resolution orbital photography of that region for site certification and surface-data (after landing) correlation.

intercept trajectories prior to their respective insertions into elliptical orbits around the planet. See Figs. 32 and 33. Timing of the Viking missions was planned to achieve the trajectory situation shown. The Mars orbit insertion maneuvers for the Vikings require significant engine burns—in the case of *Viking 1*, for example, 38 minutes, consuming 2330 pounds (1057 kilograms) of propellant. Once in the Martian vicinity, radio signals required 22 minutes in either direction between earth and Mars, thus a total of 44 minutes to execute a command and receive confirmation of that command. The general plan successfully followed in the *Viking* program involved orbiting the spacecraft in their respective orbits around Mars for several days, not only to commence imaging of the planet and certain scientific experiments, but to reconfirm earlier decisions concerning the best landing sites to finally elect for the two landers. Then, much as the *Surveyor* soft lunar landing craft had been placed on the Earth's Moon several years before, the landers were released from the aeroshells on the orbiting spacecraft, using parachute deployment at an elevation above the Martian

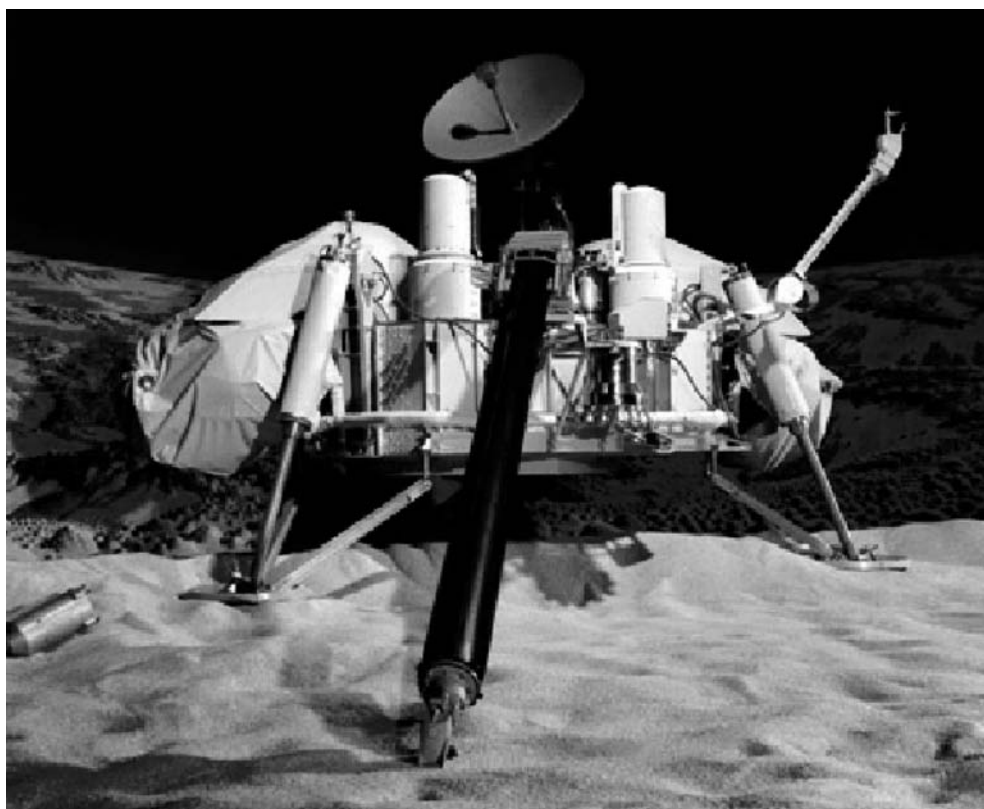


Figure 34. One of the *Viking* Landers (test model) in a simulated Martian setting. This spacecraft was set up in the auditorium at NASA's Jet Propulsion Laboratory to thoroughly familiarize the many scientists on the project with the detailed operation of the spacecraft's mechanical sampling system and other scientific experiments aboard. (NASA: *Jet Propulsion Laboratory, Pasadena, California.*) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

surface of about 19,400 feet (5,913 meters) and the firing of terminal-descent engines at about 4,600 feet (1,402 meters) above the planet's surface. Both landers touched down successfully. Later, adjustments were made to the *Viking* orbiters to provide better imaging of additional areas of the planet. Transmissions from the orbiters and the landers extending over an extensive period and information from one of the spacecraft was still being received during the early 1980s.

The Viking Landers

To assist in familiarizing many scientists at the control center in the detailed operation of the complex *Viking* landers, a test lander was installed in the auditorium at NASA's Jet Propulsion Laboratory in Pasadena, California.

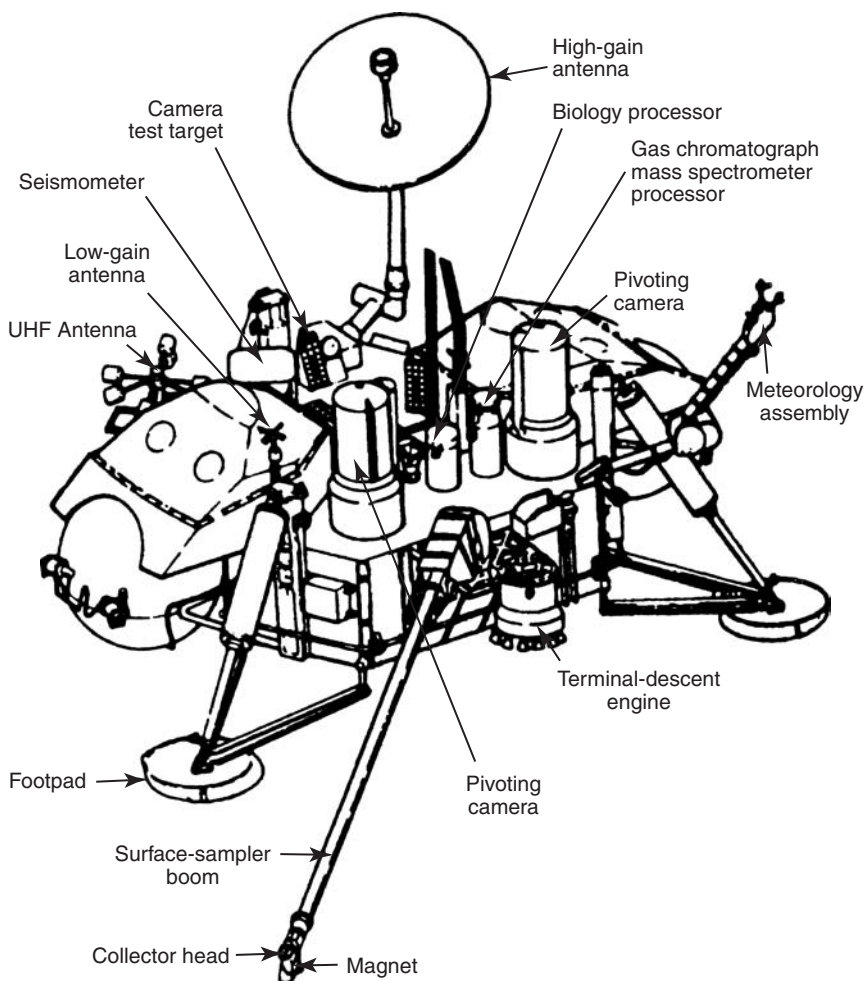


Figure 35. Principal features of the *Viking* Lander. (*Martin Marietta Corporation.*)

See Fig. 34. A diagram of one of the Landers is given in Fig. 35. Locations on Mars of the final landing sites are shown in Fig. 36. A view of the landing site for *Viking Lander 1*, taken from the orbiting *Viking 1*, prior to the landing is shown in Fig. 37.

The first photograph ever taken on the surface of Mars is shown in Fig. 38. This picture was taken just minutes after *Viking Lander 1* touched down successfully at Chryse Planitia. The center of the image is about 1.4 meters (5 feet) from camera No. 2 of the spacecraft. A similar view of the martian surface taken by *Viking Lander 2* shortly after touchdown at Utopia Planitia is shown in Fig. 39.

The first photograph of the Martian landscape (Chryse Planitia site) is shown in Fig. 40. In real color, this view is predominantly reddish brown. A diagram of the *Viking 1 Lander* and showing field of view with reference to equipment components is given in Fig. 41. Another striking view of the Martian landscape of Chryse Planitia is shown Fig. 42. The Martian landscape at the Utopia Planitia site is shown in Fig. 43.

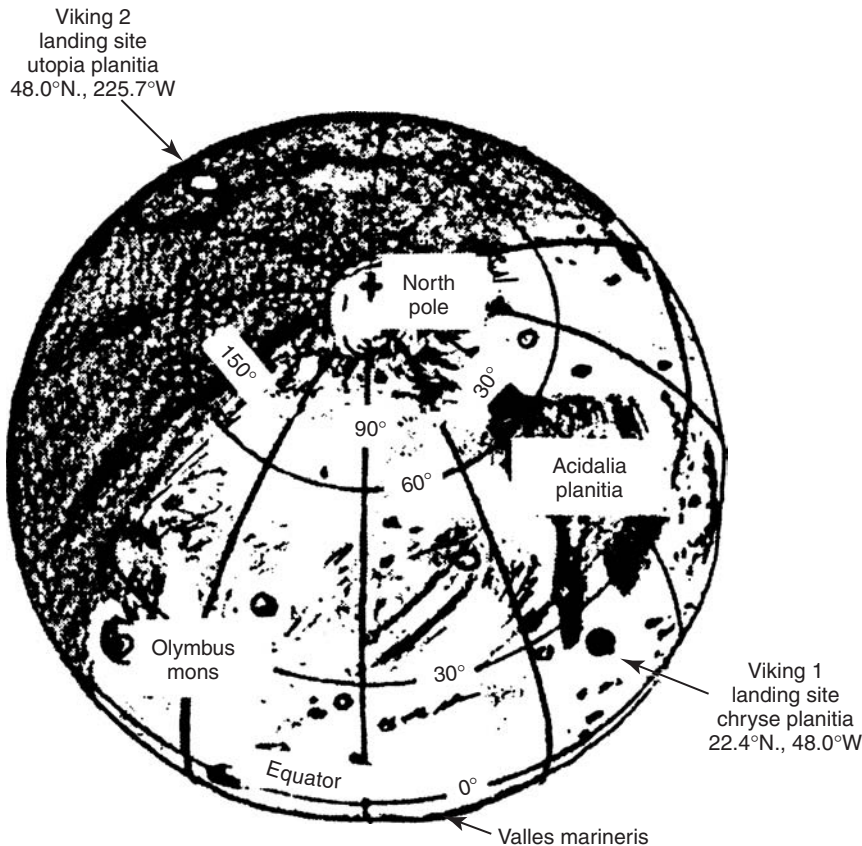


Figure 36. Locations of the two *Viking* landing sites on Mars.

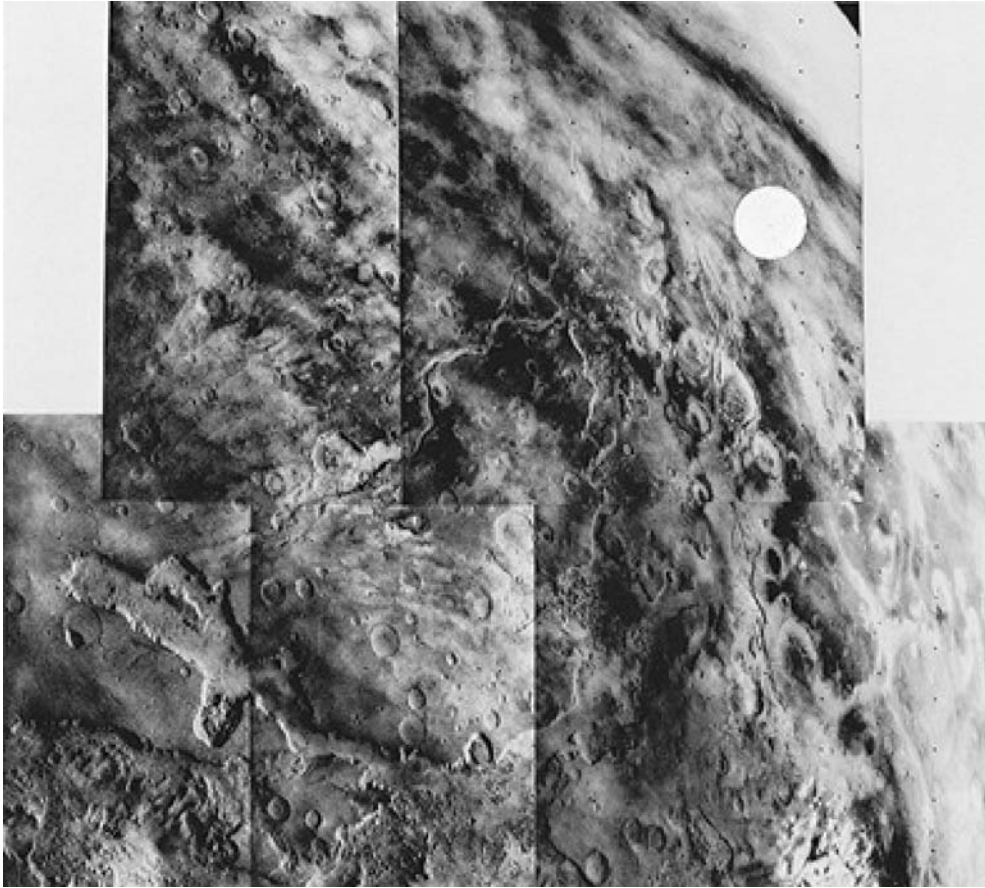


Figure 37. View from *Viking 1* Orbiter showing two candidate landing sites. White circle indicates prime site in Chryse where *Viking Lander 1* touched down a few days later. In this group of five adjoining photos taken from about 32,000 kilometers (20,000 miles) through a violet filter, Chryse is shown lying at the mouth of the channels, which proceed southward on the planet. An alternative site lies on a plateau adjacent one of the canyons in the lower (foreground) part of the picture. A bit of the planet's limb can be seen in the upper right-hand corner. Near the lower right-hand corner is a white cloud, believed to be ice crystals. From a comparison with pictures taken three minutes apart, the cloud was found to be moving about 97 kilometers (60 miles) per hour toward the upper left of view. Overall, the picture spans about 40° in longitude and 35° in latitude. The prominent feature in the lower left frame is Grangis Chasma, an arm of the great equatorial rift. North in these views is toward upper-right-hand corner. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

Viking Scientific Experiments

Thirteen scientific investigations yielded information about the atmosphere and surface of Mars. Two orbiters and landers operating for several months photographed the surface extensively from 1500 kilometers (930 miles) and directly on the surface. Measurements were made of the atmospheric composition, the

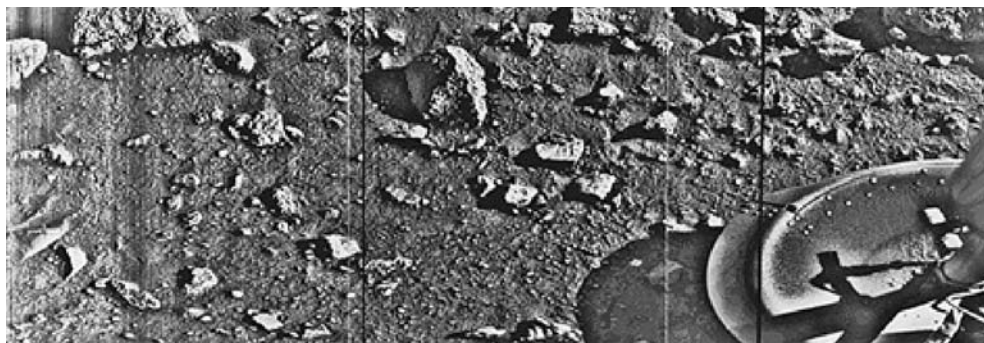


Figure 38. First photograph ever taken on surface of Mars, obtained by *Viking 1 Lander* just a few minutes after its successful touchdown on Chryse Planitia. The center of the image is about 1.4 meters (5 feet) from camera #2 of the spacecraft. Both rocks and finely granulated material—sand and dust—are observed. Many of the small foreground rocks are flat with angular facets. Several larger rocks exhibit irregular surfaces with pits and the large rock at the top left shows intersecting linear cracks. Extending from that rock toward the camera is a vertical linear dark band, which may be due to a one-minute partial obscuration of the landscape due to clouds or dust intervening between the sun and the surface. Associated with several of the rocks are apparent signs of wind transport of granular material. The large rock in the center is about 10 centimeters (4 inches) across and shows three rough facets. To its lower right is a rock near a smooth portion of the Martian surface, probably composed of very fine-grained material. It is possible that the rock was moved during the *Viking 1* descent maneuvers, revealing the finer-grained basement substratum; or that the fine-grained material had accumulated adjacent to the rock. There are numerous other furrows and depressions and places with fine-grained material elsewhere in the view. At right is a portion of footpad #2. Small quantities of fine-grained sand and dust are seen at the center of the footpad near the strut and were deposited at landing. The shadow to the left of the footpad clearly exhibits detail, due to scattering of light either from the Martian atmosphere or from the spacecraft, observable because the Martian sky scatters light into shadowed areas. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

surface, elemental abundance, the atmospheric water vapor, temperature of the surface, and meteorological conditions: direct tests were made for organic material and living organisms.⁷

Inorganic Chemistry. An X-ray fluorescence spectrometer was used to determine the elemental composition of samples at each lander site. Both sites yielded analyses of the fine-particle materials that are strikingly similar. Silica and iron in large amounts and magnesium, aluminum, calcium, and sulfur in significant amounts. More detail pertaining to the inorganic aspects of the martian surface and rock materials are given in the entry on Mars. The trenching and sampling equipment on the Lander spacecraft are shown in the foreground of Fig. 20. Plan views of the sampling apparatus and procedures followed the *Viking* Landers are given in Figs. 44 and 45.

Molecular Analysis. Two samples from each lander site were analyzed for organic material with successive use of volatilization, pyrolysis, and detection by gas chromatography-mass spectrometry (GCMS). The sensitivity of the method

⁷Information extracted from official NASA Langley Research Center report.

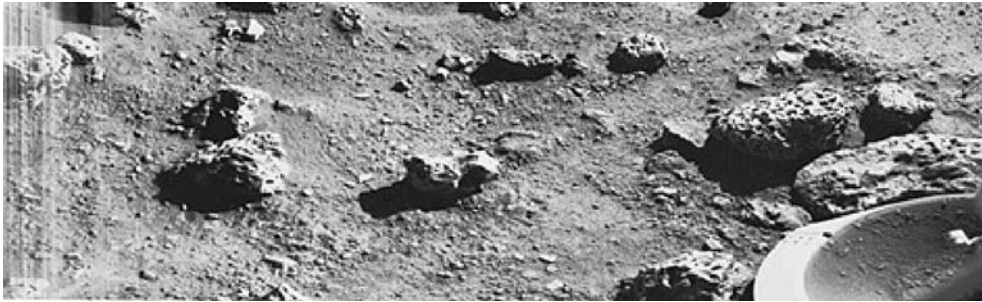


Figure 39. First photograph of Martian surface taken by *Viking Lander 2* at the Utopia Planitia site. The scene reveals a wide variety of rocks littering a surface of fine-grained deposits. Boulders in the 10–20 centimeter (4–8 inch) size range—somewhat vesicular (holes) and some apparently fluted by wind—are common. Many of the pebbles have tubular or platy shapes, suggesting that they may be derived from layered strata. The fluted boulder just above the Lander’s footpad displays a dust-covered or scraped surface, suggesting it was overturned or altered by the foot at touchdown. Brightness variations at the beginning of the picture scan (left edge) probably are due to dust settling after landing. A substantial amount of fine-grained material kicked up by the descent engines has accumulated in the concave interior of the footpad. Center of the image is about 1.4 meters (5 feet) from the camera. Field of view extends 70° from left to right and 20° from top to bottom. This second landing location is in the northern latitudes about 7500 kilometers (4600 miles) northeast of the Viking 1 Lander site, where touchdown occurred 45 days earlier. (NASA; Jet Propulsion Laboratory, Pasadena, California.)



Figure 40. First panoramic view of Martian surface taking by *Viking 1 Lander*. The out-of-focus spacecraft component toward left center is the housing for the Viking sample arm, which is not yet deployed. Parallel lines faintly seen in the sky are an artifact and are not real features. However, the change of brightness from horizon toward zenith and toward the right (west) is accurately reflected in this picture, which was taken in the late Martian afternoon. At the horizon to the left is a plateaulike prominence much brighter than the foreground material between the rocks. The horizon features are approximately three kilometers (1.8 miles) away. At left is a collection of fine-grained material reminiscent of sand dunes. The dark sinuous markings in the left foreground are of unknown origin. Some unidentified shapes can be perceived on the hilly eminence at the horizon toward left center of view. The horizontal cloud stratum can be made out halfway from the horizon to the top of the picture. At the center is seen the low-grain antenna for receipt of commands from earth. The projections on or near the horizon may represent the rims of distant impact craters. In the right foreground are color charts for Lander camera calibration, a mirror for the Viking magnetic properties experiment and pan of a grid on the top of the Lander body. At upper right is the high-gain dish antenna for direct communication between the landed spacecraft and earth. Toward the right edge is an array of smooth, fine-grained material which shows some hint of ripple structure and may be the beginning of a large dune field off to the right of the view, which joins with dunes seen at the top left in this 300° panoramic view. Some of the rocks appear to be undercut on one side and partially buried by drifting sand on the other. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

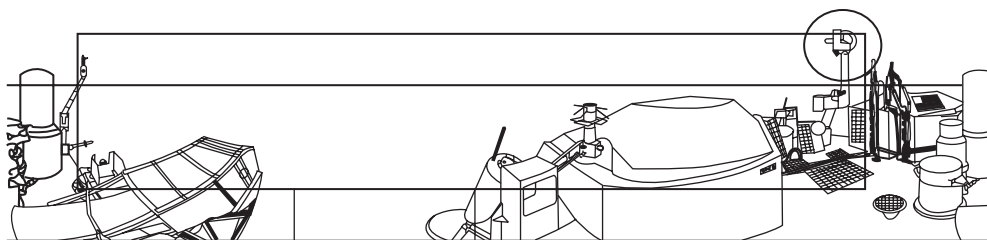


Figure 41. This diagram illustrates a full 360° image from camera #2 on the *Viking Lander* spacecraft. The outlined image areas represent the *Viking 1 Lander*'s first two pictures (Figs. 10 and 12). This type of diagram helped analysts at the mission control center to identify photo orientation in terms of pans of the Lander spacecraft components when they appeared in various views. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

is of the order of parts per billion. No organic compounds were detected at that level. The instrument was not designed to detect life—neither the quality or sensitivity permitted detecting biomass directly. The absence of organics in the sample was somewhat surprising considering the likelihood of carbonaceous chondrites reaching the Martian surface, or the possibility of *de novo* synthesis. Explanations involving dilution in the regolith and destruction by ultraviolet light or oxidation are all plausible. The GCMS was also used to measure the Martian atmosphere. It was ideally suited to measure isotopic ratios. Based upon measurements of nitrogen, argon, xenon, and krypton and their isotopic abundances, it is believed unlikely that a history of Mars outgassing will emerge.

Biology Experiments. Three experiments were conducted to test directly for life on Mars. The tests revealed a surprisingly chemically active surface—very likely oxidizing—but no evidence concerning the existence of life on the planet.



Figure 42. Martian landscape as viewed by *Viking 1 Lander*, showing a dune field with features remarkably similar to many seen in the deserts on earth. The early morning lighting (7:30 A.M. local Mars time) reveals subtle details and shading. The picture covers 100°, looking northeast at left and southeast at right. Viking scientists observed that this area is reminiscent of regions in Mexico, California, and Arizona (Kelso, Death Valley, Yuma). The sharp dune crests indicate the most recent wind storms capable of moving sand over the dunes in the general direction from upper left to lower right. Small deposits downwind of rocks also indicate this wind direction. Large boulder at left is about 8 meters (25 feet) from the spacecraft and measures about 1 × 3 meters (3 × 10 feet). The meteorology boom, which supports the spacecraft's miniature weather station, cuts through the center of the picture. The sun rose two hours earlier and is about 30° above the horizon near the center of the view. In real color, the landscape is predominantly reddish brown. (NASA; Jet Propulsion Laboratory, Pasadena, California.)



Figure 43. High-resolution photo of the Martian surface taken by Viking Lander 2 at the Utopia Planitia landing site. View was made on May 18, 1979 and relayed to earth by Orbiter 1 on June 7. The “rolling hill” Marscape is an artifact of the Lander’s 8-degree tilt; the horizon is generally flat. There is some relief of the flat terrain, however, at lower scales locally and at greater scales on the horizon. In stereo, a few gullies and depressions take on considerable depth and dimension. A thin coating of water ice on the rocks and soil is visible. The time the frost appeared corresponded almost exactly with the build-up of frost one Martian year (23 earth months) earlier when a similar view had been taken. The frost remained on the surface for about one hundred days. Some scientists believe dust panicles in the atmosphere may pick up bits of solid water. But carbon dioxide present in the Martian atmosphere also may freeze and adhere to particles, becoming sufficiently heavy to settle. Warmed by the sun, the surface evaporates the carbon dioxide and returns it to the atmosphere, leaving behind water and dust. The ice seen in this picture, like that which formed during the earlier martian year, is extremely thin, perhaps no more than 1/1000 inch (0.03 millimeter) thick. (NASA; *Jet Propulsion Laboratory, Pasadena, California.*)

The biological experiments were conducted with fully programmed and automated miniature laboratory equipment installed in each *Viking* lander. In the pyrolytic release (PR) experiment, Martian soil was placed in a chamber, after which carbon dioxide and carbon monoxide were added. These compounds were traceable because of the addition of radioactive carbon-14. The soil was incubated beneath a lamp that simulated Martian sunlight, but with no ultraviolet radiation present, as is the actual case on the planet today. If microorganisms were present, they would take up the radioactive gases. The chamber was heated to pyrolyze (decompose) any microbes present in the organic gases. The gases then were forced into an organic vapor trap, allowing other gases to pass to a radiation detector for “first count.” With additional heating, the “organic vapors” were released and, if radioactive, they would indicate that living organisms were present. Results were negative.

In the labeled release (LR) experiment radioactive nutrient was added to a soil sample. Microorganisms present would digest the nutrient and release radioactive carbon dioxide. The soil is permitted to “incubate” for a period of a week or more, with further additions of nutrient. Results were negative.

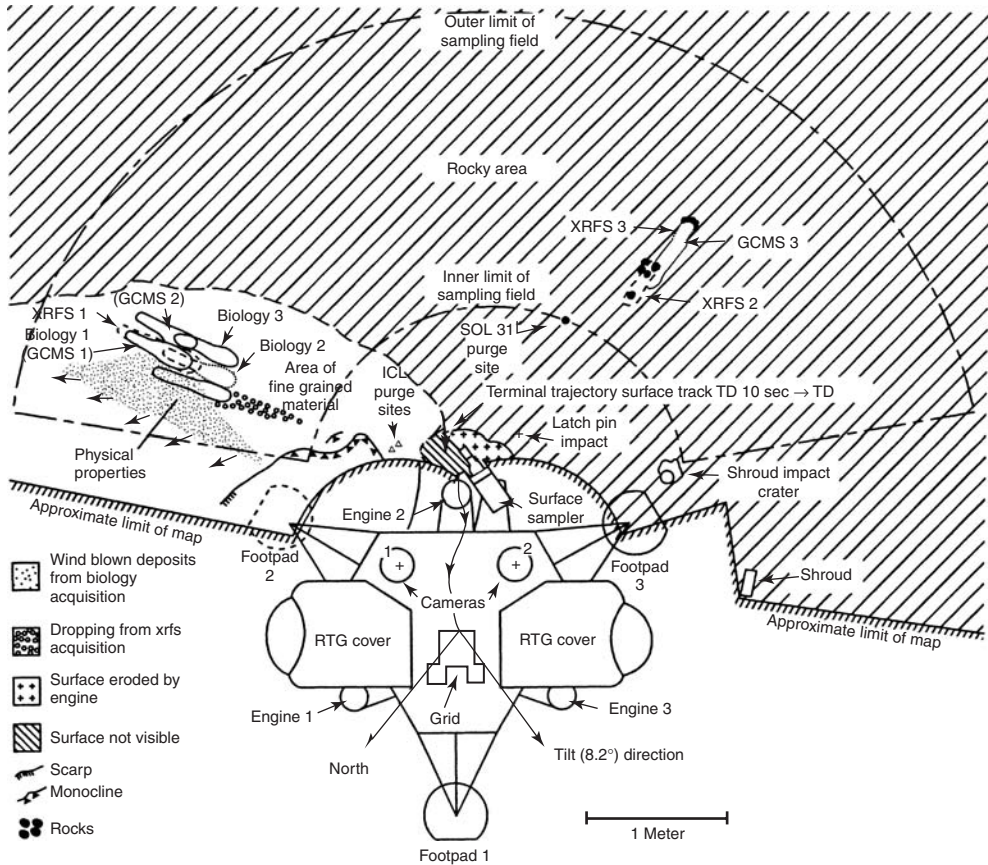


Figure 44. Plan view of *Viking Lander 1*, showing the spacecraft and its orientation, location of sample sites, locations of selected rocks for analytical experiments. 1 Sol = 1 complete Martian day and night; XRFS = x-ray fluorescence spectrometer; GCMS = gas chromatograph-mass spectrometer; ICL = initial rock to be investigated. (NASA; Jet Propulsion Laboratory, Pasadena, California.)

In the gas exchange (GEX) experiment, scientists were looking for changes that Martian microbes might cause in gas levels over a long period. Soil was placed in a chamber, which was sealed to prevent gas leakage. Just sufficient nutrient flows admixed with water vapor would awaken spores or seeds, changing the gas level in the experiment. The results were negative. Considerable production of oxygen was noted from the GEX experiment, more or less explained as unusual Martian exotic chemistry.

No plausible ties with living organisms were established. There continues to be some speculation that the chemical oxidative qualities of Martian soil may support microorganisms of a sort which earth-bound scientists have not described even on a sound theoretical basis.

Meteorology. A meteorological weather station to measure changes in pressure, temperature, wind speed and direction operated well on both landers. Generally, the weather at both sites was repetitive, with a daily temperature

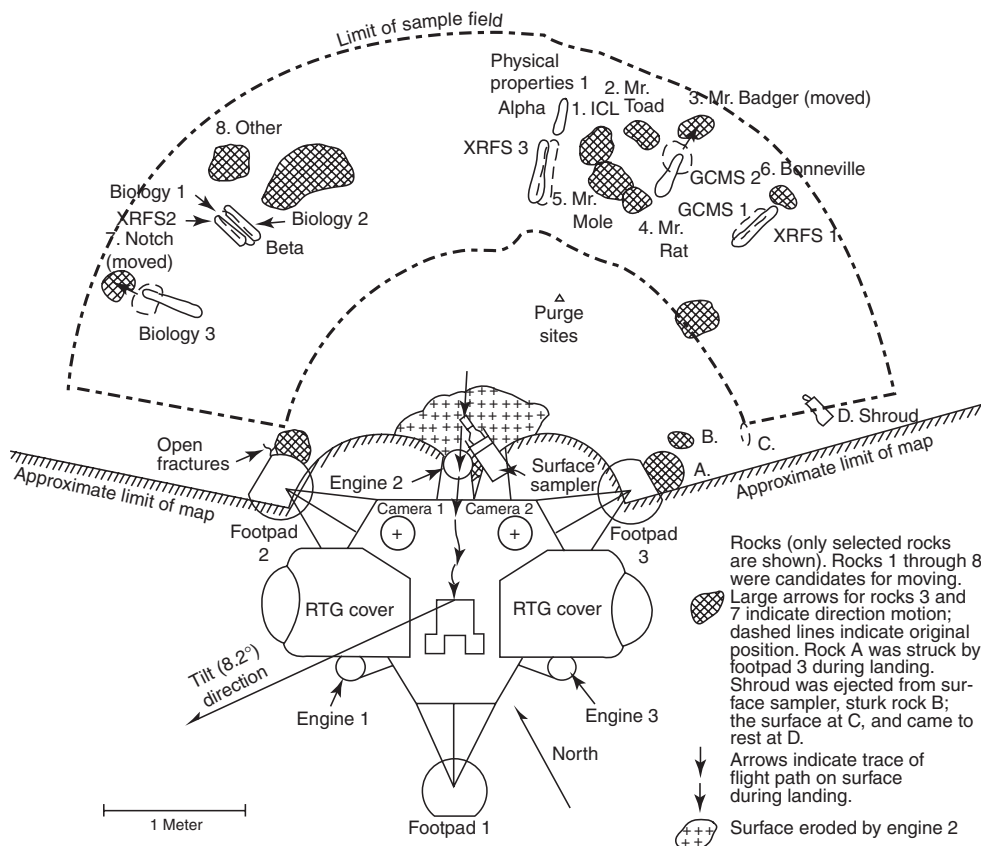


Figure 45. Plan view of Viking Lander 2, similar to that of Fig. 29. (NASA Jet Propulsion Laboratory, Pasadena, California.)

variation of between 190 and 240 K, the peak usually occurring in mid-afternoon Martian time. The pressure at each site was in the range of 7 to 8 millibars. Daily pressure variations were about 0.3 millibar.

Seismology. The seismometer on Viking Lander 1 failed to be uncaged, but the Viking 2 functioned normally. Little or no quake activity was detected.

Atmospheric Water Detector. An infrared spectrometer operating at the 1.38. micrometer region was mounted on the scan platform of each lander. The device was used to measure the latitudinal variations and diurnal variations and, by operating over a complete Martian year, it was able to indicate seasonal changes. More moisture was found in the northern hemisphere than in the southern portion of the planet. These measurements helped to confirm that the permanent polar cap of Mars consists of water ice.

Infrared Thermal Mapper. An infrared radiometer measured thermal emission of the surface and atmosphere. It was found that the atmospheric temperature above 20 kilometers (about 12 miles) varies from 165 K (near dawn) to 185 K at about 2:15 in the afternoon Martian time. This variation is believed to

be initiated at the lower levels and radiatively propagated by dust in the atmosphere. The temperature of the surface was found to be highly variable.

Physical and Magnetic Properties. Cameras were used to determine certain physical and magnetic properties of the soil. Pictures of stroke gages, sample digging, footpad movement, areas underneath the lander, and rock movement were used to determine the bulk density, particle size, angle of internal friction, cohesion, adhesion, and penetration resistance of the Martian soil. Small permanent magnets on the sampler collected material indicating that the surface contains a few percent of magnetic material, very likely magnetite.

ADDITIONAL READING (1976–1987)

1. Anders, E., and T. Owen. Mars and Earth: Origin and Abundance of Volatiles. *Science* 198: 453–465 (1977).
2. Arvidson, R.E., et al. *Three Mars Years: Viking Lander 1 Imaging Observations*. 222: 463–478 (1983).
3. Baker, V.R. The Spokane Flood Controversy and the Martian Outflow Channels. *Science* 202: 1249–1257 (1978).
4. Baker, V.R. *The Channels of Mars*. Univ. of Texas Press, Austin, Texas, 1982.
5. Bogard, D.D., and P. Johnson. Martian Gases in an Antarctic Meteorite? *Science* 221: 651–654 (1983).
6. Carr, M.H. *The Surface of Mars*. Yale University Press, New Haven, CT, 1981.
7. Carr, M.H., R.S. Saunders, R.G. Strom, and D.E. Wilhelms. *The Geology of the Terrestrial Planets*. Spec. Pubn. SP-469, National Aeronautics and Space Administration, Washington, DC, 1985.
8. Duxbury, T.C., and J. Veverka. Deimos Encounter by Viking: Preliminary Imaging Results. *Science* 201: 812–814 (1978).
9. Ezell, E.C., and L.N. Ezell. *On Mars, National Aeronautics and Space Administration*. Washington, DC, 1984.
10. Greeley, R. Release of Juvenile Water on Mars. *Science* 236: 1363–1364 (1987).
11. Haberle, R.M. The Climate of Mars. *Sci. Amer.* 54–62 (May 1986).
12. Hartman, W.K. *J. Geophys. Res.* 78: 4096 (1973).
13. LPI. Mars: Evolution of Its Climate and Atmosphere, Proceedings of Symposium, Washington, DC, Lunar and Planetary Institute, Houston, TX (July 17–19, 1986).
14. Lucchitta, B.K. Recent Mafic Volcanism on Mars. *Science* 235: 565–567 (1987).
15. McElroy, M.B., et al. Composition and Structure of the Martian Upper Atmosphere: Analysis of Results from Viking. *Science* 194: 1295–1298 (1976).
16. Mumma, M.J., et al. Discovery of Natural Gain Amplification in the 10-Micrometer Carbon Dioxide Laser Bands on Mars: A Natural Laser. *Science* 212: 45–49 (1981).
17. Mutch, R.A., et al. *The Geology of Mars*. Princeton Univ. Press, Princeton, NJ, 1977.
18. Neukum, G., and D.U. Wise. Mars: A Standard Crater Curve and Possible New Time Scale. *Science* 194: 1381–1387 (1976).
19. NEWS. Launch of Mars Observer—1992. *Science* 235: 743 (1987).
20. Owen, T., et al. The Atmosphere of Mars: Detection of Krypton and Xenon. *Science* 194: 1293–1295 (1976).
21. Paige, D.A., and A.P. Ingersoll. Annual Heat Balance of Martian Polar Caps: Viking Observations. *Science* 228: 1160–1168 (1985).
22. Pieri, D.C. Martian Valleys: Morphology, Distribution, Age, and Origin. *Science* 210: 895–897 (1980).

23. Soffen, G.A., and C.W. Snyder. The First *Viking* Mission to Mars. (contains 13 papers relating to *Viking 1 Lander* on Mars), *Science* 193: 759–815 (1976).
24. Soffen, G.A. Scientific Results of the Viking Missions. (contains 20 papers relating to Viking missions to Mars), *Science* 194: 1274–1353 (1976).
25. Soffen, G.A. Status of the *Viking* Missions. (contains 15 papers relating to the Viking mission to Mars), *Science* 194: 57–105 (1976).
26. Squyres, S.W. The History of Water on Mars. *Ann. Review of Earth and Planetary Sciences* 12: 83–106 (1984).
27. Squyres, S.W., and M.H. Carr. Geomorphic Evidence for the Distribution of Ground Ice on Mars. *Science* 231–252 (1986).
28. Thomas, P., and P.J. Gierasch. Dust Devils on Mars. *Science* 230: 175–177 (1985).

ADDITIONAL READING (1988–2000)

- Beardsley, T.M. US–Soviet Collaboration in Space Science Is Improving. *Sci. Amer.* 21 (August 1988).
- Beardsley, T. Slow Boat to Mars. *Sci. Amer.*, 14 (April 1990).
- Bergreen, L. *Voyage to Mars: NASA's Search for Life Beyond Earth*. The Putnam Publishing Group. New York, NY, 2000.
- Collins, M. Mission to Mars. *Nat'l. Geographic*. 732 (November 1988).
- Cornell, J. Red Weather (Mars). *Technology Review (MIT)*, 19 (February/March 1990).
- Eberhart, J. Soviet Findings from Phobos and Mars. *Science News*. 286 (October 28, 1989).
- Eberhart, J. Phobos: Moonlet of the Pits. *Science News*. 301 (November 4, 1989).
- Eberhart, J. Powerful Appeal of Mars' Missing Field. *Science News*. 150 (March 10, 1990).
- Eberhart, J. Episodic Oceans: Mars. *Science News*. 283 (May 5, 1990).
- Eberhart, J. The Sandy Face of Mars. *Science News*. 268 (October 27, 1990).
- Eberhart, J. Mars: Let It Snow, let it snow.... *Science News*. 286 (November 3, 1990).
- Flam, F. Swarms of Mini-Robots Set to Take on Mars Terrain. *Science*. 1621 (September 18, 1992).
- Greeley, R., and B.D. Schneid. Magma Generation on Mars: Amounts, Rates, and Comparisons with Earth, Moon, and Venus. *Science*. 996 (November 15, 1991).
- Hamilton, D.P. NASA to Explore Three Possible Mars Missions. *Science*. 863 (February 22, 1991).
- Keating, G.M.M. *Exploration of Venus and Mars Atmospheres*. Elsevier Science, New York, NY, 1995.
- Kerr, R.A. Soviet Failure at Mars a Reminder of Risks. *Science*. 26 (April 7, 1989).
- Kerr, R.A. Planetary Science Funds Cut. *Science*. 282 (January 19, 1990).
- Kieffer, H.H., C. Snyder, B.M. Jakosky, and M.S. Matthews. *Mars*. University of Arizona Press, Phoenix, AZ, 1997.
- Kiernan, V. Reactor Project Hitches onto Moon-Mars Effort. *Science*. 1482 (June 22, 1990).
- Kiernan, V. Sailing to Mars. *Technology Review (MIT)*, 20 (November/December 1990).
- McKay, C.P., and R.H. Haynes. Should We Implant Life on Mars? *Sci. Amer.* 144 (December 1990).
- Muhleman, D.O., et al. Radar Images of Mars. *Science*. 1508 (September 27, 1991).
- Ostro, S.J., et al. Radar Detection of Phobos. *Science*. 1584 (March 24, 1989).
- Owens, T., et al. Deuterium on Mars: The Abundance of HDO and the Value of D/H. *Science*. 1767 (June 24, 1988).
- Raeburn, P. *Mars: Uncovering the Secrets of the Red Planet*. National Geographic Society, Washington, DC, 2000.
- Rubincam, D.P. Mars: Change in Axial Tilt Due to Climate. *Science*. 720 (May 11, 1990).
- Schwartz, B.D. Muddy Evidence. *Sci. Amer.*, 28 (June 1989).

- Sheehan, W. *The Planet Mars: A History of Observation and Discovery*. University of Arizona Press, Phoenix, AZ, 1996.
- Staff: Mars Mission. *Technology Review (MIT)*, 19 (May/June 1989).
- Staff: Mars Magnetism: A Moot Question? *Science News*, 31 (July 14, 1990).
- Staff: Martian Atmosphere Eyed by Hubble Space Telescope. *Hughes News*, 5 (May 17, 1991).
- Staff: Instruments Help Unlock Mars' Secrets. *Hughes News*, 1 (October 2, 1992).
- Touma, J., and J. Wisdom. The Chaotic Obliquity of Mars. *Science*, 1294 (February 26, 1993).
- Waldrop, M.M. Jet Propulsion Lab Looks to Life After Voyager. *Science*, 1037 (September 8, 1989).
- Waldrop, M.M. Phobos at Mars: A Dramatic View—and Then Failure. *Science*, 1042 (September 8, 1989).
- Waldrop, M.M. Asking for the Moon. *Science*, 637 (February 9, 1990).
- Walker, M. *Evolution of Hydrothermal Ecosystems on Earth and Mars*. John Wiley & Sons, Inc., New York, NY, 1996.
- Walters, M. *The Search for Life on Mars*. Perseus Publishing, Boulder, CO, 1999.

WEB REFERENCES

- Mars Exploration: <http://Mars.jpl.nasa.gov/>
- National Aeronautics and Space Administration: http://sse.jpl.nasa.gov/missions/mars_missions/mgs.html

GLENN D. CONSIDINE
Westfield, Massachusetts

MERCURY

Introduction

Mercury, our solar system's innermost planet, is shrouded in mystery, despite the fact that it is Earth's third closest planetary neighbor and is one of the five planets known to the ancients. It is fairly easy to see. It appears to an Earth-based observer at approximately 2-month intervals as a bright object visible to the naked eye near the Sun, shortly before sunrise or after sunset. However, despite its nearness and its brightness, it ranks only behind the most distant planet, Pluto, in our understanding of its interior structure, its surface composition, and the dynamic interactions between its surface, its tenuous atmosphere, and its magnetosphere. The physical processes that have governed Mercury's origin and evolution are poorly understood. The planet has many unusual qualities, including an unexpected magnetic field and a remarkably high density (mass to volume ratio). It has even been suggested that water ice may be present just under the surface at some places on the planet, even though its surface is the closest to the Sun of all of the planets. Because Mercury is close to the Sun, the response of the planet's environment to solar activity is also expected to be greatest of all of the planets.

Mercury stands out as an object of intense scientific interest because it represents an extreme within the family of planets. It formed at the highest

temperature of all of the objects that condensed from the presolar nebula. As the closest planet to the Sun, it supports an environment that tests the limits of current theories of comparative planetology. Its surface experiences wide extremes in temperature because the planet rotates slowly in a period of 59 days and revolves about the Sun every 88 days. This 2:3 ratio of rotational period to period of revolution creates a dawn-to-dawn Mercurian day of 176 Earth days, the longest day in the solar system. The dayside equatorial temperature approaches 700 K, among the highest of any planetary surface in the solar system. When day turns to night, the temperature plunges rapidly, and the surface becomes as cold as the unilluminated side of the Moon (100 K). This is the most extreme daily temperature fluctuation of all of the planets. Understanding more about Mercury is important for testing models of planetary formation, and the results of these research activities will have great impact on models of the origin and evolution of the solar system. The basic physical properties of Mercury are shown in Table 1.

History of Exploration

The observations of the ancients did not reveal much about the surface of Mercury. Even though the planet appears bright to the naked eye, it is sufficiently far away that no surface detail can be seen. The ancients were able only to map its motion in the sky. The use of the telescope, pioneered by Galileo in 1610, did not add much to what was already known by the ancients because Mercury, as the innermost planet, is never more than 27° from the Sun as seen projected on the celestial sphere from the perspective of an Earth-based observer. From Earth, the planet can either be observed during the day, when scattered sunlight provides a strong source of background noise, or low in the sky while contaminated by background signal from Earth's atmosphere when the Sun is just over the horizon, shortly before sunrise or after sunset. In this case, as Mercury approaches the horizon, the light must pass through 5–10 times as much turbulent air as when it is observed overhead. This limits the

Table 1. **Physical Properties of Mercury^a**

Mass	3.301×10^{23} kg
Mass relative to Earth	0.0554
Semimajor axis	0.387099×10^6 km
Eccentricity	0.205628
Inclination to ecliptic	$7^\circ 0' 15''$
Radius	2425 km
Bulk density	5.44 g/cc
Sidereal period	87.969 days
Period of rotation	58.646 days
Surface gravitational acceleration	363 cm/s^2
Escape velocity	4.2 km/s
Dipole magnetic moment	$2-6 \times 10^{12} \text{ Tm}^3$

^aFrom Reference 1.

ability of ground-based astronomers to observe Mercury. Thus, the typical spatial resolution obtainable on Mercury even with the best Earth-based telescopes is just a few hundred km, far worse than can be seen on the Moon by the unaided eye.

The modern generation of space-based observatories, pioneered principally by the International Ultraviolet Explorer and currently dominated by the Hubble Space Telescope, are not limited by the atmospheric distortion problems of observing Mercury from Earth. However, both of these space-based observation platforms are constrained from pointing near the Sun due to the concerns of spacecraft engineers that direct rays of the Sun might pass down the telescope and damage the sensitive optical instruments. Given these difficult geometric observing constraints, it is difficult to enhance vastly our understanding of Mercury using available observing techniques from Earth's surface or from near-Earth orbit.

Despite these problems, terrestrial observing has produced some interesting results. For example, in 1955, astronomers reflected radar waves from Mercury's surface. By measuring the Doppler shift in the frequency of the returned radiation, it was established that Mercury has a 59-day rotational period, two-thirds that of its 88-day orbital period of revolution. Until then, it had been thought that Mercury had an 88-day rotational period, identical to its period of revolution about the Sun. It had been assumed that Mercury was in synchronous rotation about the Sun just as the Moon is about Earth. This finding proved to be just one prelude to the many surprises from Mariner 10, two decades later (2).

The Mariner 10 Mission

Mariner 10, the Mercury-Venus flyby mission, elevated our understanding of Mercury from almost nothing to most of what we presently know. Getting Mariner to Mercury was not a trivial task. The planet's orbit lies deep in the gravitational potential well of the Sun, and therefore Mariner flew by Venus to exchange gravitational energy and thus slow the spacecraft for a Mercury encounter. The heliocentric orbit provided three close flybys of Mercury on 29 March 1974, 21 September 1974, and 16 March 1975. The spacecraft carried an ensemble of instruments intended to address fundamental questions about the planet's current physical state, its origin, and subsequent evolution.

The Mariner imaging system was based on the most current television technology available when the mission was conceived. The spacecraft returned to Earth approximately 2000 images of about 45% of the planet. The resolution of the images was about 1.5 km, comparable to that of the Moon when viewed from Earth through a telescope. The other side of Mercury, more than half the planet, has never been seen. Despite Mariner's accomplishments, our present understanding of Mercury remains limited. We know as much about Mercury today as we knew about our own Moon at the dawn of the space program (3,4).

Mercury's Unusual Density

Mariner results confirmed that one of Mercury's most unusual distinguishing characteristics is its remarkably high bulk density, the planet's mass divided by its volume. It was found that Mercury's density is 5.44 grams/cc, remarkably greater than its apparent surficial likeness, the Moon. It has been known since the time of Archimedes that density differences measured between objects imply that the objects have different physical states and/or chemical compositions. Except for Mercury, the terrestrial planets (Venus, Moon, Mars, and Earth) all exhibit a fairly linear relationship between density and size. Figure 1 illustrates this point. The largest terrestrial planets, Earth and Venus, have the highest density; the smaller ones, Moon and Mars, have densities that are much lower. However, Mercury is unique. It is relatively small, comparable to the Moon and Mars, but its density is typical of that expected from a much larger planet such as Earth or Venus. This density "anomaly" is an important gateway to understanding Mercury's interior.

Planets, in general, do not have homogeneous interiors. The outer layers are usually composed of low-density materials such as silicate rocks whose density is about 3 gm/cc. However, the density of a planetary interior increases with depth due to the compression created by the overlying rock layers and gravitational fractionation of the materials that comprise the interior. This causes the

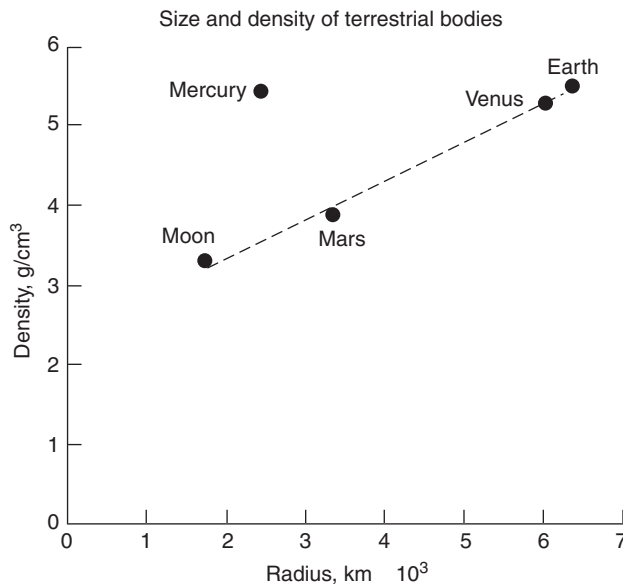


Figure 1. It has been known since the time of Archimedes that the density of an object yields important information about the object's internal constitution. Archimedes used this principle to determine the gold content in the crown of a monarch, but the density of Mercury shows that it is unique among the terrestrial planets. The other terrestrial planets conform to a linear relationship between size and density. Mercury, however, is relatively small, yet its density is comparable to that of large objects like Earth and Venus. This has been understood to mean that Mercury has a very large core comprised mostly of iron, a very dense material, and a relatively thin crust of lower density silicates.

densest material to become concentrated at the center. Based on our understanding of the cosmic abundance of the elements, it is generally accepted that the high density materials that comprise the cores of the planets are extremely rich in iron and may even contain pure iron in some cases. Density considerations strongly suggest that Mercury has the largest core relative to planetary size of all of the terrestrial planets. This fact has stimulated a lively debate regarding the processes that governed the origin and evolution of the solar system.

The unique internal structure of Mercury is a fundamental constraint on models of its origin and evolution, and these models, in turn, constrain the models of solar system formation. Mercury's bulk density (5.44 g/cc), which rivals that of Earth (5.52 g/cc) and Venus (5.25 g/cc), is furthermore unusual when the effect of the relatively larger masses of Earth and Venus are considered. The larger objects have greater overburdens that compress the interiors of Earth and Venus more than Mercury's smaller exterior mass compresses its interior. When this compressional effect is considered, Mercury, it is found, has the largest uncompressed density (5.3 g/cc) of all of the terrestrial planets. (Earth's uncompressed density is about 4.4 g/cc) (5). This can best be explained by assuming that Mercury has the largest fraction of iron of all of the terrestrial planets and the Moon. Mercury's core, where the iron is concentrated, is about three-fourths the size of the entire planet. For comparison, Earth's core is about half its total size.

The current hypothesis of solar system formation holds that all of the planets condensed from the solar nebula at about the same time. If this is true then, one of three circumstances may have occurred that made Mercury so different from the other terrestrial planets. First, the composition of the solar nebula might have changed dramatically from the vicinity of Mercury's orbit to the regions of the other terrestrial planets, and this change may have been much larger than theoretical models would predict; second, early in the lifetime of the solar system, the Sun's energy was so great that a large fraction of the more volatile, low-density elements on Mercury were vaporized and driven off; or third, Mercury suffered a catastrophic episode, such as a collision with a very massive object some time after it formed, which drove off the less dense materials. Another catastrophic possibility is that Mercury formed elsewhere in the solar system and was moved to its present location by a random act of violence. All of these very different hypotheses can be entertained, given the current body of evidence. Future missions to Mercury are expected to help resolve these issues.

Mercury's Strong Magnetic Field. Mariner 10 made three close passes by Mercury; two of were at altitudes of 300 and 700 km. The magnetometer on the spacecraft measured magnetic field strengths of 100–400 nT. This yields a dipole magnetic moment of $\sim 4 \times 10^{12} \text{ Tm}^3$. Mercury's dipole magnetic field is among the largest of those of all of the terrestrial planets, except Earth. This field is represented by the lines shown in Fig. 2. It is generally believed that Earth's magnetic field is supported by electric generator-like processes that occur in the liquid circulating in Earth's interior and that a field is produced as a consequence of the relative motions between the conducting regions in the interior. If Mercury's magnetic field has a similar origin, then this would imply that the planet

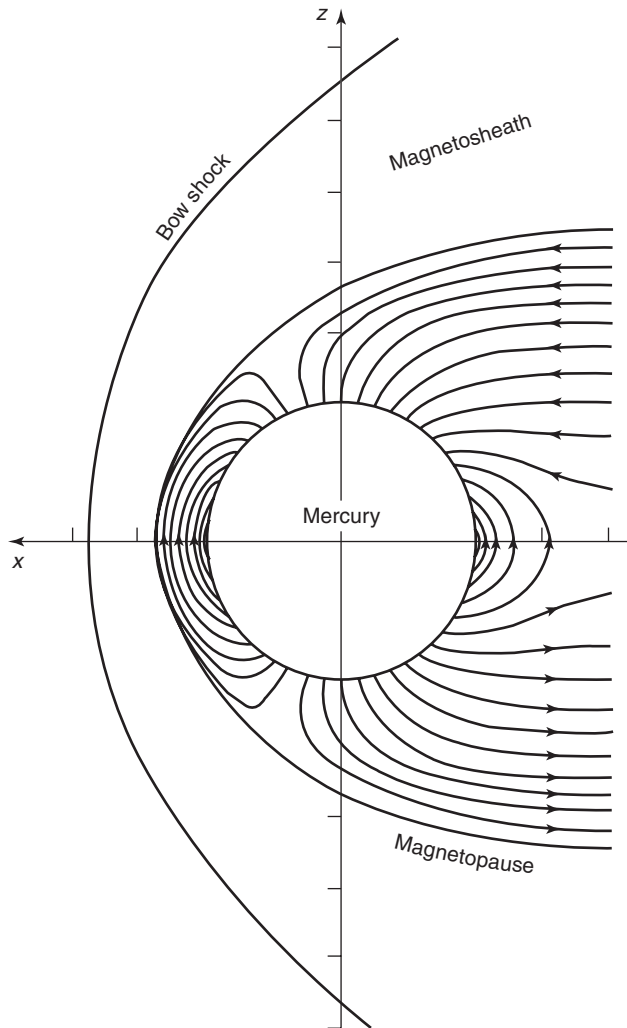


Figure 2. Mercury has a relatively strong dipole magnetic field. Earth is the only terrestrial planet with a larger field. Dipole magnetic fields, it is believed, originate from the motion of liquid materials in a planet's interior. The presence of the field implies that Mercury has a partially molten core. This is surprising because objects as small as Mercury would be expected to have solidified aeons ago. The magnetic field traps electrically charged particles about Mercury forming a magnetosphere. The magnetosphere is constantly being deformed by the solar wind, a flow of particles streaming out from the Sun. At times, the solar wind becomes so intense, that it pierces the magnetic field and directly touches the planet's surface. This creates a constantly changing physical environment for the material comprising the surface.

has a liquid core (6). However, small objects like Mercury have a higher proportion of surface area compared to volume than the larger planets and therefore, other factors being equal, small bodies dissipate their energy to space faster. If Mercury has an iron core, as the density and magnetic field data imply, then a pure iron core should have solidified aeons ago. Yet, a solid core could not support

a self-sustaining dynamo to create a magnetic field. In response to this apparent contradiction, it has been proposed that other materials are probably present in the core and, along with iron, these additional materials form a eutectic with the iron which depresses the freezing point, just as antifreeze depresses the freezing point of water in an automobile radiator. Sulfur, a cosmically abundant element, has been suggested as a possible candidate which, when combined with iron, would keep Mercury's core liquid at a much lower temperature than would be expected from a pure iron core alone. If so, then the dynamo process could sustain a magnetic field (7).

Mercury's Surface Composition. The high abundance of iron in Mercury's interior is in sharp contrast to the apparent dearth of iron in the planet's crust. Careful photometric analysis of the Mariner data along with laborious spectroscopic telescope observations from Earth have failed to detect even trace amounts of iron in Mercury's crustal rocks. On the other hand, iron has been detected by spectroscopic observations of the surfaces of the Moon and Mars, and, it is known, to be present in the crustal rocks of Earth. This implies that Mercury is one of the most differentiated objects in the solar system, with all of the high-density materials such as iron being concentrated in the interior, and only the low-density silicates being present in the crust.

Iron is evident in the reflection spectrum of lunar rocks by an absorption feature at 0.95 microns. This feature has also been measured in the spectrum of returned lunar samples measured in the laboratory. It is also seen in the spectra of iron oxides of terrestrial origin. There have been reports in the literature that this 0.95-micron absorption feature has been seen on Mercury; however, it has been argued that all of the spectra that report this feature are compromised by the incomplete removal of features caused by the terrestrial atmosphere, which also absorbs in the same spectral region. To date, no positive uncompromised evidence exists for this spectral feature. Observations at microwave wavelengths (7) and mid infrared spectroscopy (8) have been interpreted as inconsistent with iron being present in the crust in a fashion similar to the lunar crust. If Mercury does not have iron-bearing minerals in its crust, then the obvious interpretation of this result is that the iron on Mercury is very strongly concentrated in the planet's interior rather than on the surface (10).

Alternatively, Mercury's surface is often exposed to the highly reducing effect of bombardment by the solar wind. This reduces iron oxides to pure metallic iron which does not exhibit the spectral feature at 0.95 microns.

The intensity of light reflected from a typical planetary surface changes as the angle of illumination (incidence) and observation (emission) change. The rate of this change depends on the surface composition and texture. Although a comprehensive photometric study of Mercury's surface awaits a Mercury orbiter mission, the limited Mariner data set finds little to distinguish Mercury's photometric properties from those of the Moon (11). Normal reflectances derived for different geologic regimes on Mercury are between 0.09 and 0.45. The range in reflectances observed in lunar materials is similar and ranges from 0.05–0.40 in some areas. However, the dark areas of the Moon are darker than the darkest areas of Mercury. Given the vast differences in the quality of the data sets used for photometric analyses of both bodies, these differences may not be of major significance.

However, Mercury has yet to be studied at a wide range of different angles of illumination and emission; this task will fall to future orbiter missions that will measure the scattering properties of Mercury's surface from very small to very large phase angles.

If it is found after a full photometric study is completed, that the surfaces of the Moon and Mercury have similar photometric properties, the result might still not be considered unusual. Even if the Moon and Mercury evolved along very different paths, the evolution of the surfaces of both bodies may have been driven by similar processes in recent epochs, creating an evolutionary convergence imposed by similar environmental circumstance.

Mercury's Geologic History

A cursory glance at an image of the surface of Mercury shows a remarkably lunar-like morphology. The surface is dominated by craters; the largest is Caloris, 1300 km in diameter. The crater density increases with decreasing size to the lowest limit of the Mariner 10 cameras.

The floors of the larger crater basins are less cratered than the immediately surrounding terrain, consistent with the hypothesis that these largest basins are younger in age. The surface is also marked with lineament features and a unique set of thrust faults, or lobate scarps, that may be due to a global change in shape which occurred after the surface had solidified.

Careful inspection of the Mariner 10 images of Mercury's surface found no evidence for currently active volcanism. It is generally believed that large-scale volcanic processes on Mercury's surface ended with the heavy cratering episode that concluded the solar system's planetary formation phase which ended about 4 billion years ago. More recent lava flows on Mercury have occurred on only a small scale, as shown in Fig. 3 and are associated with impacts rather than internal processes (12).

Once planetary surfaces solidify, they no longer exhibit substantive viscous behavior. Instead, when subjected to external forces, the surface either responds elastically or else it fractures as a brittle solid. A planetary surface, that behaves this way is called a lithosphere. The paucity of volcanic activity on Mercury has preserved a record of the earliest tectonic events that occurred shortly after its brittle lithosphere developed. This preservation of tectonic events that occurred early in Mercury's history is a record that was obliterated or overprinted and obscured on the larger more active planets.

Mercury has several large craters that are surrounded by multiringed material of which Caloris is the largest. This 1300-km diameter behemoth it is estimated, was formed 3.6 billion years ago. It serves as a Mercurian geologic reference indicator because its effects are widespread and define a relative single reference point in time. The antipode of Caloris is characterized by hilly and lineated terrain, suggesting that the impact was so catastrophic that disruption of the surface may have occurred on the side of Mercury opposite the impact.

Mercury's surface is crosscut by rectangular features of unknown origin, that are preferentially oriented north-south, northeast-southwest, and northwest-southeast. These lineaments are called the Mercurian grid. A probable

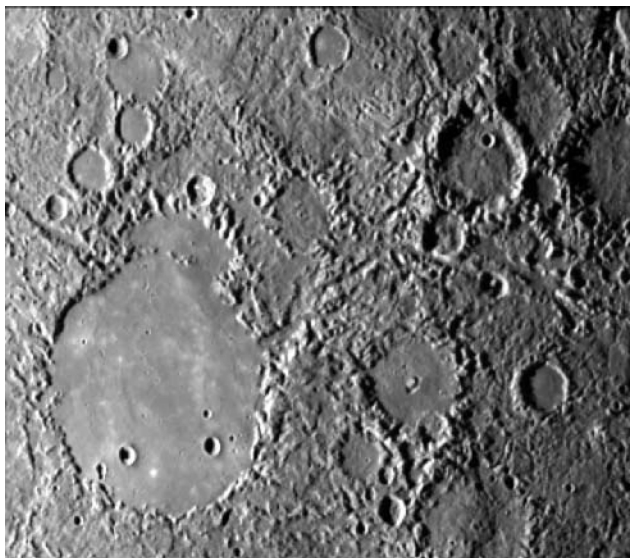


Figure 3. Petrach crater is a relatively recent addition to Mercury's surface. This is evidenced by the relative paucity of craters, compared to the surrounding terrain. The event that created Petrach may have melted so much rock that lava may have flowed through a channel and flooded the crater about 100 km away.

explanation of the lineament pattern is that Mercury's crust solidified at a time when the planet was rotating much faster, perhaps only about 20 hours. The model posits that the crust would have solidified while the planet was rapidly spinning and the forces of hydrostatic equilibrium would have created an equatorial bulge, characteristic of a rapidly rotating, self-gravitating object. After the planet slowed to its present rotational period, gravity pulled it into a more spherical shape. The lineaments it is thought, were created as the lithosphere accommodated the new gravitational environment. It has been noted that the lineaments do not cut across the huge Caloris crater, indicating that the Caloris event occurred after the lineament structure had been established (see Fig. 4).

While Mercury was rotationally slowing, it was also cooling. It is thought that at least part of the core solidified. The core shrinkage that accompanied this solidification probably resulted in a net decrease of about a million square kilometers in surface area. This decline in lithospheric surface area would have produced the network of thrust faults that are expressed as a series of lobate scarps that crisscross Mercury's surface. It has also been noted that the network of lobate scarps has a fortuitous orientation with respect to Caloris and may be an expression of lithospheric response to such a catastrophic impact. See Figs. 5 and 6.

These depictions of Mercury's geologic history have been constructed in the absence of accurate age estimates for the geologic units that comprise the surface. The only way in which the absolute age of a rock can be determined is by radiometric dating of returned samples. Geologists, however, have numerous ingenious ways of determining the relative ages of surface units, mostly based on the principle of superposition, which, simply stated, is that a feature that overlies

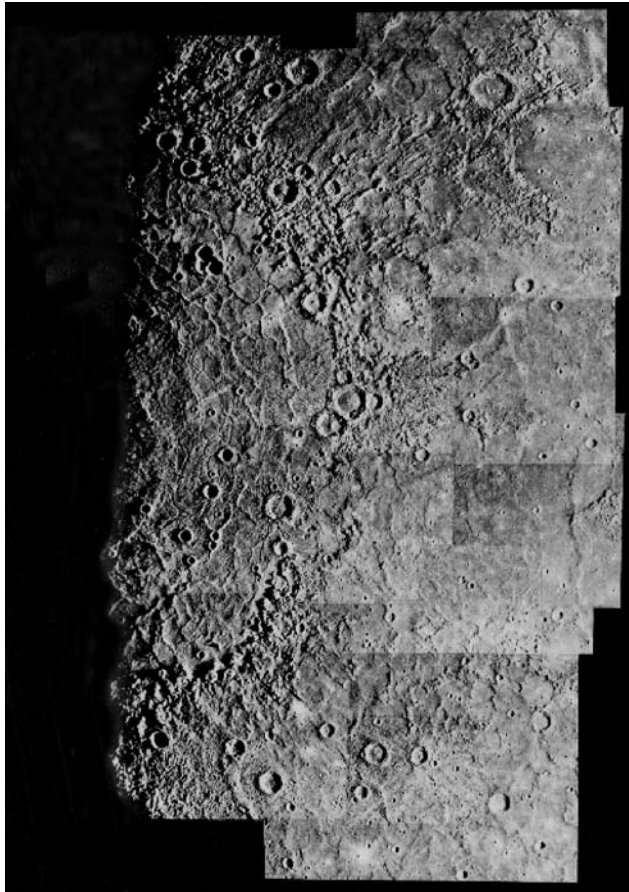


Figure 4. The Caloris basin is one of the largest craters on any object in the solar system. It spans 1300 km in diameter. Cratering statistics indicate that the Caloris impact was relatively recent in Mercury's history, not part of the heavy bombardment that occurred early in the history of the solar system. The reason for this is that Caloris' features cover the lineated grid across much of Mercury's surface. It is believed that the grid pattern resulted from cracking of Mercury's lithosphere due to slowing of the planet's rotation early in its history.

or cuts across another is the younger of the two features. This principle is particularly helpful in determining the relative ages of craters.

Throughout the solar system, planets, planetary satellites, and other smaller solid surface bodies show widespread evidence of impacts from still smaller objects. Even our own planet Earth, where the constant effects of erosion and the slow drift of the tectonic plates tends to erase surface features, shows many craters on its surface. The objects that create these craters typically strike the surface at velocities in the neighborhood of 10 km/s, fast enough to vaporize the impacting object along with a significant amount of planetary material.

The final shape of a crater is related to the physical properties of the impactor, its velocity, angle of impact, and the properties of the material in which the crater is excavated. To first order, the size of a crater is related to the energy of

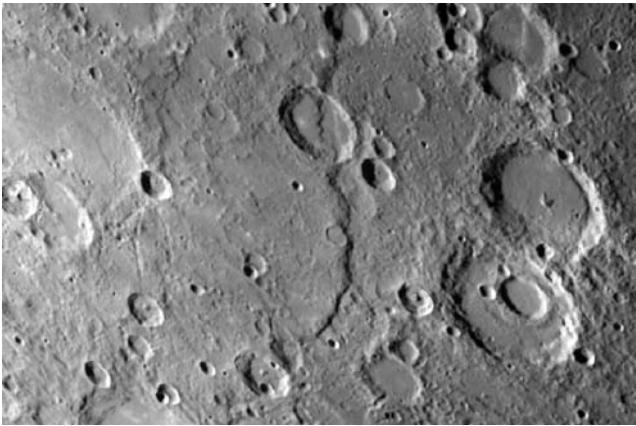


Figure 5. Discovery scarp is typical of the scarps on Mercury. The scarps are younger than many of the craters that are on Mercury’s surface. Therefore, it is observed that they cut across them. These scarps it is believed, are thrust faults that were created by crustal shrinking of the planet as the core partially solidified.

the impact. Kinetic energy is a function of mass and velocity. Therefore, the larger and faster impactors cause the largest craters. Study of the important crater features such as size, depth and height of the central peak, slope of the rim walls, provides information about conditions at the time of the impact.

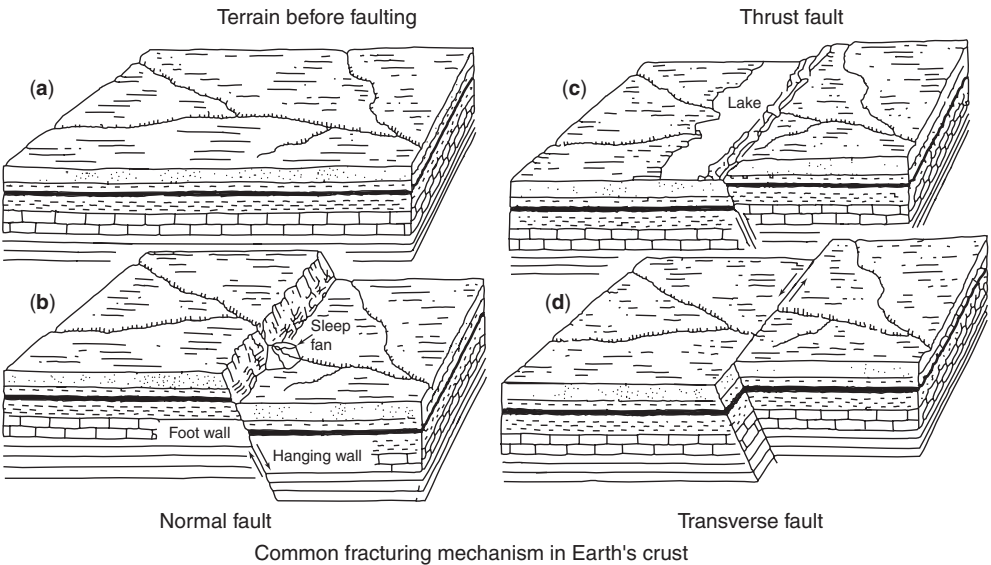


Figure 6. The motion of the relative blocks of crustal material is illustrated in the attached schematic of the common faulting mechanisms on Earth. (a) The unbroken terrain prior to faulting. (b) A simple or normal fault caused by the surface stretching or pulling apart. (c) A thrust fault that is created when the crust is compressed. (d) A transverse fault typical of those found on Earth where continental plates are moving laterally to each other. Mercury’s surface is riddled with thrust faults, which indicates that its crust shrank relatively late in its history.

The size–frequency distribution of craters is a most useful measure for a comparative planetology assessment. The heavily cratered highland surfaces of Mercury, Mars, and the Moon are the product of the period of late heavy bombardment period in solar system history. The three planets have similar crater size–frequency distributions but the relative sizes of the Mercurian craters are larger indicating that the impacting objects had a higher velocity when they struck Mercury. This is expected, given that the laws of kinematics dictate that objects in gravitationally bound, elliptical orbits move about the Sun move faster when they are near the Sun. Therefore, when these objects impact Mercury, which is closer to the Sun than the other planets, they have a relatively higher velocity and hence they create a larger crater size relative to the impacting object's size. This is consistent with the hypothesis that these impacts were from the same family of objects in orbit about the Sun. The dissimilarity of the crater populations seen on the surfaces of bodies in the inner solar system and the outer solar system indicates that the objects that impacted the inner planets were from a distinct parent group that did not extend beyond Mars to the outer solar system.

Mercury's Tenuous Atmosphere

Planetary atmospheres are described as thick or thin on the basis of the density of the atmospheric gases and their temperatures. The important discriminator between a thick atmosphere and a thin atmosphere is the mean free path of the atmospheric molecules. The mean free path is the mean distance that the average particle travels before colliding with another particle. Planets that are as hot as Mercury do not retain appreciable atmospheres because the thermal motions of the low molecular weight volatiles exceed the escape velocity from the planet. The density of the remaining gases is so low that the individual molecules move on ballistic trajectories. The probability of collisions between individual molecules is small. By analogy, high above the surface of Earth, the density of atmospheric gases becomes so low that the few atmospheric particles present move similarly on ballistic trajectories. The part of an atmosphere that meets this condition is called the exosphere. On Earth, the exosphere begins about 500 km above the surface (13). On Mercury, the entire atmosphere is the exosphere.

Because of the high temperatures, it was expected that any significant amount of volatile material on Mercury would soon be lost to space, and hence, it had long been thought that the planet did not have an atmosphere at all. The atmospheric gases are hot and therefore the particles move at high velocity. The particles that comprise Mercury's atmosphere travel in ballistic trajectories and their courses are altered only by Mercury's gravity and collisions with its surface.

However, the ultraviolet spectrometer on Mariner 10 did detect small amounts of hydrogen, helium, and oxygen, and subsequent Earth-based observations have detected small amounts of sodium and potassium (14). The source and ultimate fate of this atmospheric material is a subject of lively discussion within the scientific community (see Fig. 7). Much of the atmosphere is probably

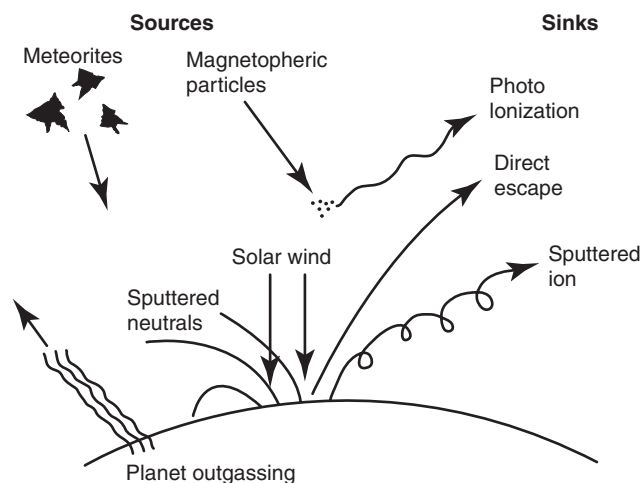


Figure 7. Surprisingly, Mercury has a tenuous atmosphere. This is not expected on a surface that is so hot. The sources and sinks of the atmospheric material is the subject of a lively debate within the scientific community. Possible sources could be slow outgassing from the planet itself. Furthermore, the solar wind probably plays an important role in creating Mercury's atmosphere as the solar wind particles strike the surface and sputter material into the atmosphere. Cometary and meteoric material may be falling on the surface in sufficient quantity that their small volatile components may be supplied to the atmosphere. The atmosphere may escape by a variety of processes. Hot particles have velocities high enough that they can escape the planet's gravitational field directly. It is also possible that they may be hit by energetic particles from Mercury's magnetosphere, which charges and captures them in the magnetic field.

due to the action of thermal evaporation and to the effect of the solar wind, a stream of energetic particles, which on occasion directly bombards Mercury's illuminated side. The solar wind is 10 times greater at Mercury's orbit than it is at Earth. These particles can remove material from Mercury's surface by the process of sputtering. Once an atom is sputtered off the surface it becomes part of Mercury's tenuous atmosphere. In addition, some of Mercury's atmosphere may originate from materials recycled from the Mercurian magnetosphere or by direct infall of meteoric or cometary material. It is also possible that the planet is directly outgassing the last remains of its primordial volatile inventory.

Recent radar observations from Earth have suggested that the radar reflective properties of material near Mercury's poles are similar to those of water ice (15). The prospect that a planet as close to the Sun as Mercury has ice caps (or any water at all) is quite intriguing. It has been argued that the ice may reside in permanently shaded regions near Mercury's poles and may be the remains of primordial water that condensed on Mercury when the planet formed from the solar nebula. If the source of this ice is indeed primordial water, then Mercury must have remained in a remarkably stable orbit for the entire age of the solar system, never tilting either pole toward the Sun, despite catastrophic events such as the Caloris impact. This would be a highly remarkable circumstance.

Another possible source of water might be a rain of cometary material that is falling into Mercury from all directions. Material that falls at the pole may

remain in the shade and may be only slowly eroded. The proposed polar water deposits may be a source of Mercury's atmospheric oxygen and hydrogen. For example, it is possible that the shaded polar regions may contain another volatile species whose radar reflectivity is similar to that of water ice. The radar observations may be explained by the presence of volatile materials other than water. Sulfur, a cosmically abundant element, has been suggested as a possible volatile, that might explain the radar observations. Nevertheless, the suggestion of volatile condensates near Mercury's poles implies a dynamic atmosphere driven by processes that are, at present, poorly understood.

Mercury's atmospheric material may experience several diverse fates. The high-velocity particles may be lost to space directly by thermal escape, the slower moving particles may reimpact Mercury's surface and become redeposited, or the neutral atoms may lose one or more electrons and become ionized and then leave the planet as a consequence of electromagnetic forces generated by the planet's magnetic field and any electric fields present.

The Mercurian Magnetosphere

Mercury's magnetic field is strong enough to trap charged particles about Mercury forming a magnetosphere that is a miniaturized version of Earth's. When Mercury is at aphelion, far from the Sun, the magnetic field has sufficient strength to prevent the solar wind from reaching the planet's surface. It is estimated that the solar wind standoff distance is between 1.1 and 3 Mercury radii. Magnetospheres constantly change in response to external conditions, most notably the activity of the Sun. However, the timescale for magnetospheric changes is much shorter for Mercury's smaller magnetosphere than for Earth's. Thus, combined studies of the two magnetospheres can give information on the response time of a planetary magnetosphere to perturbations.

In Earth's case, the thick atmosphere provides the lower boundary of magnetospheric activity. Mercury, however, lacks a thick atmosphere; hence, the lower limit is provided by the surface of the planet itself. In fact, unlike Earth, a large fraction of Mercury's magnetosphere is occupied by the solid surface of the planet. As a consequence, Mercury's magnetosphere lacks several important features that play a significant role in Earth's magnetosphere. This includes the inner radiation belts and the plasmasphere. The energetic particles in the Mercurian environment are principally the solar wind and from additional lesser contributions of cosmic rays and Jovian electrons.

The boundary between the solar wind plasma and the magnetospheric plasma is called the magnetopause. This boundary regulates the amount of solar wind energy that is deposited on the planet's surface. As the Mariner 10 spacecraft passed through the magnetopause, the spacecraft magnetometer and ion plasma analyzers changed. The magnetic field strength and orientation changed as the pass-through occurred.

Mercury's magnetic field is just strong enough to prevent the solar wind from reaching the planet's surface except when solar activity is high or when Mercury is at perihelion, the point in its orbit that takes it closest to the Sun. In these instances, the solar wind reaches the surface directly. When this happens,

the solar wind particles sputter particles directly from Mercury’s surface to its atmosphere and magnetosphere. Thus, the analysis of the magnetosphere is important for understanding the planet’s atmospheric and surface composition. It is impossible to measure the variation in the intensity of Mercury’s magnetosphere from Earth (14).

Obstacles to Exploration of Mercury

In the four decades that humans have had access to space, the combined efforts of the world’s spacefaring nations have produced only one spacecraft visit to Mercury. This occurred more than two decades ago in 1973 when NASA sent the Mariner 10 spacecraft to explore Venus and Mercury, the two planets of the inner solar system. Since then, other nearby objects have enjoyed the attention of the space programs of the United States, the former Soviet Union, the European Space Agency, and Japan, but Mercury has been ignored. From the dawn of the space age to the turn of the millennium, the United States sent more than 40 missions to explore the Moon, 20 to Venus, and more than 15 to Mars. By the end of the next decade, an armada of orbiting spacecraft will have been placed in orbit about Venus, Mars, Jupiter, and Saturn that will return detailed information about these planets and their environs on timescales of years. Mercury however, will still remain largely unexplored, despite its many mysteries. Figure 8 illustrates the number of missions that have been flown to the nearby planets and the Moon.

There are several explanations for this skewed emphasis. Exploring these questions provides interesting insights into the way the United States, the world’s leading space power, manages its responsibility to explore the solar system.

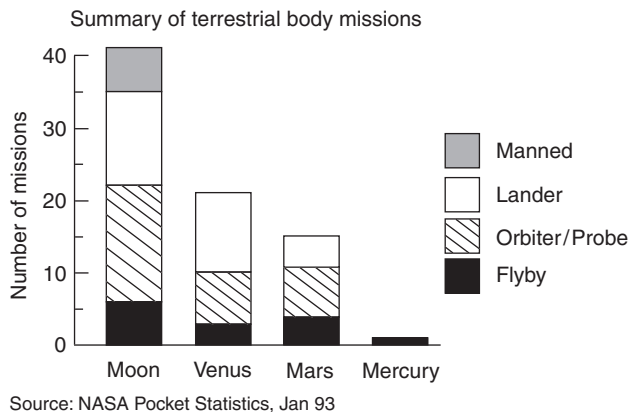


Figure 8. Summary of missions that have been flown to explore the terrestrial planets. Mercury is noteworthy in that it has been explored only once more than 20 years ago by the Mariner 10 spacecraft. This disparity in emphasis has occurred in spite of NASA’s commitment to a balanced program of space exploration where relatively equal emphasis is placed on exploring all of the planets.

A prominent obstacle to further Mercury research lies in the popular, but untrue, belief that Mercury is merely another lunar-like object. Ironically, this is in part due to the success of Mariner 10. That spacecraft returned television images of about 40% of Mercury's surface, and they showed, to first order, that Mercury has a heavily cratered surface similar to that of the Moon. However, surface morphology is one of the very few similarities between Mercury and the Moon. These two objects have grossly differing physical attributes that suggest very different evolutionary histories. Unfortunately, because of the tremendous importance attached by human consciousness to visual images, there remains a widespread misconception that Mercury and the Moon are nearly identical objects, which by happenstance, occupy different regions of the solar system.

Another major hurdle to further exploration is technical. A spacecraft in orbit about Mercury must be protected, first, from the intense energy from the Sun, and second, as the spacecraft nears Mercury's sunlit side, from the solar energy reflected from Mercury. At those times when the spacecraft is between the Sun and Mercury, the energy reflected from the planet's surface can become a greater threat to the spacecraft than the Sun itself. It is flying between a frying pan and a fire. Maintaining spacecraft thermal stability in this situation is a major problem. The addition of thermal mitigation techniques increases the spacecraft mass and stresses the limits of the available launch vehicle technology, particularly if the spacecraft is to undertake its interplanetary cruise using conventional chemical bipropellants. However, the recent success of NASA's New Millennium Deep Space One mission in demonstrating solar electric ion propulsion systems for spacecraft in deep space has demonstrated that alternative spacecraft propulsion techniques, such as ion drives, are feasible, can reduce the launch mass of a Mercury probe significantly, and can permit the mass of the science payload to increase. The next exploration of Mercury by the United States will still rely on conventional chemical bipropellant techniques, and this mission may provide limited answers to some of the questions that baffle us about Mercury. This will be immediately followed by a Mercury mission sponsored by the European Space Agency. This mission will be based on more modern technology. The solar electric propelled ESA mission will provide a significant advance in our knowledge of Mercury and will permit the scientific community to undertake intense comparative studies of Mercury and the other planets in the solar system. Such solar electric powered spacecraft can be expected to dominate the exploration of the inner solar system and to provide platforms to enhance our knowledge of difficult outposts such as Mercury.

BIBLIOGRAPHY

1. Allen, C.W. (ed.). *Astrophysical Quantities*, 3rd ed. Athlone Press, London, 1976; see also Connerney, J.E.P., and N.F. Ness. Mercury's magnetic field and interior. In F. Vilas, C.R. Chapman, and M.S. Matthew (eds). *Mercury*. Univ. of Arizona Press, Tucson, 1988, pp. 494–513.
2. Nelson, R.M. Mercury: The Forgotten Planet. *Sci. Am.* (November 1997).
3. Vilas, F., C. Chapman, and M.S. Matthews (eds). *Mercury*. University of Arizona Press, Tucson, 1988.

4. Beatty, J.K., and A. Chaikin (eds). *The New Solar System*. Cambridge University Press and Sky Publishing, Cambridge, U.K. and Cambridge, MA, 1990.
5. Ringwood, A.E. *The Origin Of The Earth And The Moon*. Springer-Verlag, New York, 1979.
6. Ness, N.F. The magnetic fields of mercury, Mars, Moon. *Ann. Rev. Earth Planetary Sci.* 7: 249–288 (1979).
7. Stevenson, S. In *Mercury*, Vilas et al., op.cit.
8. Mitchell, D.L., and I. De Pater. Microwave imaging of Mercury's thermal emission at wavelengths from 0.3 To 20.5 cm. *Icarus* 110: 2–32 (1994).
9. Sprague et al. Mercury's Feldspar Connection: Mid-IP Measurements Suggest Plagioclase. *Adv. Space Res.* 19: 1507–1510 (1997).
10. Vilas, F. Absence of crystalline Fe^{2+} in the regolith. *Icarus* 64: 133–138 (1985).
11. Hapke, B.W. Interpretation of Optical Observations of Mercury and the Moon. *Phys. Planet. Inst.* 15: 264–274 (1977).
12. Davies, M.E., D.E. Gault, S.E. Dwornik, and R.G. Strom (eds). *Atlas of Mercury*. NASA Scientific and Technical Information Office, Washington, DC, 1978.
13. Shirley, J.E. Exosphere. In *Encyclopedia of Planetary Science*. Chapman and Hall, New York, 1997, p. 248.
14. Potter, D., and T. Morgan, Potassium in the atmosphere of mercury. *Icarus* 67: 336–340 (1986).
15. Slade, M., B. Butler, and D. Muhleman. Mercury: Radar Imaging For Polar Ice. *Science* 258: 635–640 (1992).
16. Russell, C.T. Flux Transfer Events At Mercury. *J. Geophys. Res.* 90: 11067–11074 (1985).

ROBERT M. NELSON
Jet Propulsion Laboratory
Pasadena, California

MILITARY GROUND CONTROL CENTERS, UNITED STATES

Military Ground Control Centers

Most people have the perspective of military operations centers typified by the 1980s movie “War Games” (1). They expect to visit facilities such as Colorado’s Cheyenne Mountain Complex (2) and see vast, darkened underground chambers ringed with immense display screens supported by powerful computers and occupied by legions of console operators. Although an attempt was made in the mid-1990s to “spruce up” Cheyenne Mountain Complex and other military space facilities, the fact is that military operations centers are mostly filled with rather antiquated equipment and present a decidedly unmodern appearance. Anyone who has flown on a military aircraft would understand the less than spectacular ambience and would feel right at home in the military’s space operations centers.

Current military space operations centers grew out of 1950s and early 1960s facilities and missions—some of those missions were quite different from

those of today. The military is quite conservative when it comes to innovation and modernization. And nowhere is this more evident than in space operations. The watchword is evolutionary, not revolutionary. Thus, a facility, and even its hardware, has often been adapted from a previous mission in the same facility. Because the old mission, often critical to national security, must continue as new missions are brought on, old hardware and procedures are replaced only when new technologies and hardware are well proven. Similarly, organizations responsible for military space missions tend to be adapted from older, often quite different missions and organizations. In recent years, this is particularly evident in the U.S. Air Force's space organizations that are steadily being evolved out of and in a way that mimics air operations. Organizations such as the Air Force Space Command's 50th Space Wing, the operator of much of the Service's on-orbit satellites, is organized exactly as a flying wing, complete with the assumption of a unit name. As this article unfolds, the reader should keep this conservatism and legacy in mind.

The U.S. military currently maintains ground control centers for three related missions: early warning, space surveillance/space control, and on-orbit satellite operations. Early warning and space surveillance/space control evolved from the North American Air Defense mission developed jointly by the United States and Canada in the 1950s, culminating in the construction of the North American Air Defense Command (NORAD) Combat Operations Center (COC) in the early 1960s (2). On-orbit satellite operations centers have a different pedigree. Shrouded in secrecy during most of the Cold War, these facilities began as control centers for the United States space-based intelligence collection mission. As such, they were originally developed and largely operated by the still secretive National Reconnaissance Office (NRO) (3). To understand current military operations centers, it is necessary to understand the history of these missions and organizations.

History

North American Air Defense Command (NORAD). NORAD is the consolidated United States–Canada organization responsible for the defense of North America (2). Its key command and control center is contained in the Cheyenne Mountain Complex south of Colorado Springs, Colorado. NORAD originated in the early 1950s as a response to the development of the Soviet atomic bomb in 1949 and their TU-4 bomber designed to deliver these weapons. By the mid-1950s, the Soviet Union had thermonuclear weapons and a jet-powered bomber. In 1954, the United States established a Continental Air Defense Command with a concrete-block Command Operations Center (COC) at Ent Air Force Base in Colorado Springs, Colorado. When the Soviet intercontinental ballistic missile (ICBM) was demonstrated in the late 1950s, the perceived levels of threat in both the United States and Canada escalated and the Ent AFB above-ground facilities were considered extremely vulnerable to attack. In 1958, the U.S. and Canadian governments formally established NORAD to meet this air and missile threat. It was soon decided to build a much more survivable NORAD Combat Operations Center (COC). By the early 1960s, the Cheyenne Mountain site south of Colorado

Springs had been selected for an underground NORAD COC, and construction began in 1961 (2). Ironically, by the time the facility was completed in 1965, ICBM accuracy had progressed to the point where it is doubtful that the Cheyenne Mountain facility could survive any concerted nuclear attack (Fig. 1).

By the mid-1960s, the NORAD COC was fully operational in a series of buildings more than 1000 feet underground. The facility was designed to withstand the effects of a nuclear detonation and operate for more than a month without outside contact. Its initial mission was to use complex computer systems (which have been upgraded many times since the mid-1960s) to collect and assess data from early warning radar sensors in the United States, Canada, and other locations for potential air or missile attack on North America. By the addition of satellite missile warning sensors in the early 1970s, this data was augmented with space-based missile attack sensors.

By the mid-1960s, growing concern over Soviet Union space activities led the United States Government to include an integrated Space Defense Center as part of NORAD and as a third mission (with air and missile attack) for the NORAD COC. The Space Defense Center's function was and remains today to identify and track all man-made objects in space using inputs from missile warning radars and a dedicated set of space-track radars and optical space-track sensors.

In the mid-1980s, the United States embarked on its Strategic Defense Initiative (SDI) to develop technologies that could be applied to an effective national and global missile defense system. This initiative was highly controversial and had as its core many sensor and weapon concepts, which would be based in space. The Canadian Government had expressed in the 1980s and continued to express into the 2000s considerable concern with the arms control implications of missile defense and military activities in space. As a consequence, the role of NORAD as joint United States–Canada organizations in either missile defense or



Figure 1. The Cheyenne Mountain Operations Center is the heart of the missile warning and space track functions that support NORAD and the U.S. SPACE COMMAND, respectively. Figure courtesy USAF Space Command. Picture can be found at <http://www.spacecom.mil/images/cmdctr-4.gif>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

what is now known as space control is uncertain. Thus, most U.S. space control activities are now concentrated in a separate United States Space Command (USSPACECOM).

United States Space Command

To understand the current United States command and control of space capabilities, one must understand the basic structure of U.S. military command. U.S. military operations are carried out through a “unified” command structure, as mandated by the Goldwater–Nichols Act of 1986 (4). Responsibility for combat or other operations devolve from the U.S. President through his Secretary of Defense to the Joint Chiefs of Staff (JCS). Composed of the Service Chiefs of the U.S. Army, Navy, Marines, and Air Force, as well as a Chairman and Vice-Chairman, the JCS control U.S. military forces through a series of Unified Commands, led by Combatant Commanders. Most Unified Commands are “regional” in that they have responsibility for geographic regions such as the Pacific and Indian Ocean rims—in this case the responsibility of the U.S. Pacific Command or “PACOM.” There are also three “global commands”—the U.S. Strategic Command, responsible for nuclear war planning, the U.S. Transportation Command, responsible for global mobility, and the U.S. Space Command (USSPACECOM) responsible for space capability support for other commands, space control, and recently computer network defense and attack. The Commander of the U.S. Space Command controls and conducts space operations through forces provided by the military services. The U.S. Strategic Command and U.S. Space Command are scheduled to merge on 1 October 2002 (5).

The U.S. military services (Air Force, Army, Navy, Marines) are not responsible for military operations. Rather they train, organize and equip forces, which they provide to the various unified commanders. Thus, now there exists a U.S. Army, U.S. Navy, and U.S. Air Force Space Command to provide CINCSPACE with trained organized and equipped space forces.

The intensive focus on U.S. President Reagan’s Strategic Defense Initiative (SDI) for missile defense announced on 23 March 1983 led the U.S. Department of Defense to establish a Unified Space Command on 23 September 1985. As noted before, the Canadian government had serious reservations about the U.S. SDI and associated issues. Thus, NORAD and the USSPACECOM remain separate entities. However, since its inception, the USSPACECOM and NORAD have been dually commanded by a single U.S. Air Force four-star general. This results in a rather complex command relationship and numerous opportunities for confusion as to roles and missions. In addition, from 1990 to 2002, this Air Force general has been “triple-hatted” as the Commander of the U.S. Air Force Space Command. This arrangement again changed with the assignment of a separate four-star Air Force General to command the Air Force Space Command on 19 April 2002. This reorganization was a result of the findings of the Space Commission Study completed in 2001 (6).

A new Combatant Command, the United States Northern Command, will be initiated in late 2002. It will assume responsibility for homeland defense of the United States. Its Commander will probably retain “dual hat” Command of this

U.S. Northern Command and NORAD—unless the two Commands are formally merged by international agreement between Canada and the United States (7). At the same time, as noted before, the space and information operations responsibilities of the U.S. Space Command will devolve onto a separate Combatant Commander who will assume these and current responsibilities of the nuclear war-fighting U.S. Strategic Command.

U.S. Air Force, Army, and Naval Space Commands. Throughout the history of the U.S. military space program, the U.S. Air Force has remained a central player. Thus, its composition for training, organizing, and equipping military space forces must perforce initiate any discussion of Service programs. In the 1960s, the Air Force's Systems Command, then the research, development, and test and evaluation command, developed space systems. More discussion of this Command and its activities will follow. Because most space capabilities during the 1960s were "experimental," the operational Air Force played a little role in space system operations. The notable exception was NORAD's U.S. component, the Continental Air Defense Command (CONAD) which operated the space surveillance and missile warning system from its NORAD Combat Operations Center in Cheyenne Mountain, Colorado (8). The functions of CONAD were later assigned to the Air Force's newly named ADCOM, and CONAD was disestablished (9).

In the 1970s, NORAD's air defense mission had withered to less than 25% of its levels in the 1960s. As a result, ADCOM was disestablished as a major Air Force command in 1980. Because the Air Force traditionally assigned space systems functionally to the command or agency with the greatest need, space capabilities had, by the late 1970s, become scattered to many different organizations. Air Force Systems Command retained control of communications satellites, and the Strategic Air Command managed meteorological satellite products. However, the late 1970s saw a resurgence in military space interest. Spearheaded by the Secretary of the Air Force, Hans Mark, a variety of efforts were kicked off in the late 1970s to bring focus to the diverse Air Force space efforts. A major impetus for this restructuring was the promise of NASA's Space Shuttle as a revolutionary, and then assumed, sole future space access system and the imminent deployment of the Global Positioning System (GPS). The pressure accelerated in the early 1980s when President Ronald Reagan focused on a defense and corresponding space buildup. Thus, the Air Force formed its Air Force Space Command in September 1982 to consolidate its operational efforts. Initially, this command was linked with the Systems Command's development center ("Space Division") through a "dual-hatted" three-star general officer who served as Air Force Space Division Commander and Air Force Space Command Vice-Commander.

Today, the Air Force Space Command (AFSPC) operates most of the nation's military space infrastructure, controlling through its operations centers the space tracking and space surveillance systems, meteorological satellites, many military communications satellites, early warning satellites, and navigation satellites (GPS). It also acquired operations of the nation's ICBMs in the late 1980s. AFSPC has more than 35,000 personnel and "owns" more than 60 operational satellites along with the nation's military space launch capabilities and ranges, and satellite control network (10). However, it does not operate a national

reconnaissance system—that remains the responsibility of the National Reconnaissance Office (NRO) to be discussed later. Unlike the Air Force, the U.S. Army developed its current space organization through its long-standing missile defense mission. The Army's current space command was created as the U.S. Army Space and Missile Defense Command (SMDC) on 1 October 1997 (11). Although the Command began in 1947 when the Army created the first program office for ballistic missile defense, SMDC today amalgamates its missile defense research and development functions and its space development and operations activities. Unlike the Air Force, both development and operational programs are contained in the same organization. Key to the Army's space efforts was its pioneering effort to apply national intelligence capabilities managed by the NRO to tactical uses. Since 1973, the Army Space Program Office has overseen the Army's "Tactical Exploitation of National Capabilities" or "TENCAP." Army Space Command (ARSPACE), a subordinate command of SMDC, serves today as the Army's operational component of the U.S. Space Command (analogous to the Air Force Space Command). It does operational space planning and oversees the Defense Satellite Communications System (DSCS) Operations Centers (ground facilities for the central wideband military communications system; the Air Force Space Command operates the satellites through its satellite operations facility at Schriever Air Force Base, Colorado). The Army Space Command also explores the feasibility of off-the-shelf technology in the space program, such as the lightweight GPS receiver used during Operation Desert Shield/Desert Storm in 1991–1992.

The U.S. Navy also established a Naval Space Command on 1 October 1983 (12). It is the naval component of the U.S. Space Command. It has about 300 military and civilian personnel at the Naval Surface Warfare Center in Dahlgren, Virginia, and it operates the Naval Space Surveillance Center. Along with the Air Force Space Surveillance Network and other inputs received from missile tracking radars, it provides data to the U.S. Space Command's Space Control Center in Cheyenne Mountain, Colorado, from which the United States derives its current space situational awareness. The Naval Space Command also operates the Navy's communications satellite systems, the Fleet Satellite Communications System (FLTSATCOM) and its new Ultra-High-Frequency Follow-On ("UFO") system. The Naval Space Command operates the alternate space control center, a backup for the U.S. Space Command's Space Control Center (SCC) in Cheyenne Mountain, Colorado Springs, Colorado. The U.S. Navy announced a reorganization in 2002; its Space Command will be subsumed under a new Naval Network Warfare Command (13).

USAF Space and Missile Systems Center (SMC). The Space and Missile Systems Center traces its ancestry back to the Western Development Division (WDD). WDD was activated in July 1954 and was redesignated as the Air Force Ballistic Missile Division (AFBMD) in June 1957 (14). Its initial Commander was General Bernard A. Schriever—essentially the "father" of the U.S. military space program. The original mission of the organization was to develop intercontinental ballistic missiles (ICBMs) for the Air Force, but responsibility for developing the first military satellite system was added in February 1956. The ICBM mission remained with AFBMD and its successors through the decades that followed, but the Department of Defense (DOD) reassigned the space mission

several times before settling on a final pattern. In February 1958, the DOD activated the Advanced Research Projects Agency (ARPA) and placed it in charge of all DOD space programs during their research and development phases. In September 1959, ARPA lost its dominant role, and the DOD divided responsibility for developing military satellites among the three services. The Army was to develop communication satellites; the Navy, navigation satellites; and the Air Force (i.e., AFBMD), reconnaissance and surveillance satellites. The Air Force was also to develop and launch all military space boosters. This arrangement continued until March 1961, when the DOD gave the Air Force (AFBMD) a near monopoly on the development of all military space systems, ending the role of the Army and the Navy except under exceptional circumstances. The final policy change occurred in September 1970. The DOD declared that the Air Force would remain responsible for developing, producing, and launching space boosters and for developing, producing, and deploying satellite systems for missile warning and for surveillance of enemy nuclear delivery capabilities. However, all three military departments would have the right to submit proposals for development of satellite systems for other purposes, and DOD would decide whether to approve those proposals. In theory, this decision gave considerable leeway to the Army and the Navy and eroded the Air Force space monopoly to a considerable degree. In practice, however, the Air Force has continued to develop most of the satellite systems used by the DOD. This arrangement has undergone further change again as a result of the 2001 Space Commission Study.

As the importance of space systems increased, space and missile functions were separated in April 1961, when AFBMD was inactivated and replaced by the Ballistic Systems Division (BSD) and the Space Systems Division (SSD). In July 1967, the space and missile functions were reconsolidated to save money in the Space and Missile Systems Organization (SAMSO). Space and missile functions were separated again in October 1979, when SAMSO was divided into the Space Division and the Ballistic Missile Office. In March 1989, SD and BMO were renamed Space Systems Division (SSD) and Ballistic Systems Division (BSD). Due to the end of the Cold War and the subsequent demise of the Soviet Union, there was a significant decrease in mission programs. Consequently, BSD (which again was renamed the Ballistic Missile Organization) was realigned under SSD. Finally, in July 1992, SSD was renamed the Space and Missile Systems Center (SMC), and BMO was formally inactivated in September 1993; thus space and missile development went full circle in four decades back to a single organization responsible for both space and missile programs (14).

As the organizational structure of SMC has changed, so has the structure of its field units. Beginning in the 1950s, SMC's predecessors acquired units that controlled DOD satellites in orbit, conducted satellite and R&D missile launches, and operated the launch ranges on the East and West Coasts. The satellite control function was originally performed by the 6594th Test Group, created in 1959, and later performed by the Air Force Satellite Control Facility, which replaced the Test Group in 1965. Test wings at both Vandenberg AFB, California, and Cape Canaveral AFS, Florida, performed launch activities. Then, to reduce costs, President Carter directed that the eastern and western launch ranges be redesignated the Eastern and Western Space and Missile Centers (ESMC and WSMC) under the control of the Space and Missile Test Organization (SAMTO), created

in 1979. When the Air Force Space Command (AFSPC) was created, responsibility for operational command of military space systems passed from the test and development community to an operational organization. This resulted in the dissolution of SAMTO, and responsibility for the ESMC and WSMC was transferred to the new operational command. Even though the responsibility for space launch and on-orbit satellite operations diminished, SMC remains an important part of the development and acquisition of all military space systems. This importance was recently restated when the 2001 Space Commission Study recommended the transfer of SMC from the Air Force's Material Command to the Air Force Space Command (6).

National Reconnaissance Office (NRO)

In the late 1950s, considerable creative turmoil occurred in the organization of U.S. national security space programs. As a result of this turmoil, late in 1960, the Department of Defense created the National Reconnaissance Office (NRO) to work satellite reconnaissance programs. Although nominally attached to the Air Force, usually through its Director holding a senior Air Force civilian position (e.g., Undersecretary of the Air Force), the NRO is part of the intelligence community and not under direct control of the Air Force and its leadership.

In 1994, the NRO's existence, as well as details and data from the initial CORONA program, were declassified. The NRO continues to develop and operate the nation's reconnaissance satellites, still under classified conditions and largely separate from other national security systems.

Missions

Today's military use of space is divided into four main areas: (1) space force enhancement—the use of space capabilities to support terrestrial land, maritime, and air operations; (2) space operations—the necessary capabilities and systems to launch, operate, and, if necessary, deorbit military space systems; (3) space control—systems and functions to obtain space situational awareness (space surveillance), protect friendly use of space assets, deny enemy use of these assets and if, necessary, negate hostile space systems and capabilities; and (4) space force application—using systems that fly through space (ICBMs) or operate in space (in the future such systems as the Space-Based Laser, or SBL, under development by the USAF) to deliver lethal force to land, maritime, or air targets. For the purpose of this article, ground operations centers are considered part of the second function, space operations. But these centers support space force enhancement and space control. Organizationally, these operations centers may be further divided into those for early warning of missile attack; those for “tactical” space force enhancements systems such as the communications, navigation, intelligence and reconnaissance, and weather; and finally, those for the space surveillance portion of space control.

Space Force Enhancement: Missile Warning. The current missile warning system—the Defense Support Program (DSP) and its follow-on, the Space-Based InfraRed System (SBIRS), grew out of one of the three original 1959

satellite programs of the Air Force's Western Development Division's Military Satellite System—the Missile Detection Alarm System (MIDAS). MIDAS was aimed at developing a satellite that would carry an infrared sensor to detect hostile ICBM launches. After an initial failure, the first MIDAS satellite was successfully launched in May 1960 (14). The current Defense Support Program (DSP) began operations in the early 1970s. It consists of a number of (DSP) satellites in geosynchronous Earth orbit (GEO), ground operations centers, and missile warning centers. DSP provides a critical part of current “dual phenomenology” requirements to provide at least two unambiguous, independent ways to assess an attack against North America (Fig. 2). The other means are NORAD's ground-based early warning radars.

DSP is operated a little differently from other U.S. military satellite systems in that its satellite bus operation can be completely separate from the sensor mission operations. The 50th Space Wing's 1st Satellite Operations Squadron at Schriever Air Force Base, Colorado, can perform the former command and



Figure 2. The Air Force Space Command-operated Defense Support Program (DSP) satellites are a key part of North America's early warning systems. Figure courtesy USAF Space Command. Picture can be found at <http://www.peterson.af.mil/hqafspc/library/FactSheets/FactSheets.asp?FactChoice=6>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

control support. This function uses the Air Force Satellite Control Network (AFSCN) for transmitting command and control information and is essentially global. The latter, which can include both satellite operations and warning sensor functions, is done through dedicated operations centers and antennae. The primary site for this function is the 21st Space Wing's 2nd Satellite Warning Squadron at Buckley Air National Guard Base, near Denver, Colorado. Routine operations use the dedicated systems. But when a satellite is in critical stages of operations or maneuvers, the more flexible, global, AFSCN system is used.

The missile warning system is undergoing a major upgrade, which will enable much better missile warning, as well as make a comprehensive theater and national missile defense feasible. The Space Based Infrared System (SBIRS) is a planned constellation of high- and low-altitude satellites that has a consolidated, common ground system built to meet U.S. surveillance needs through the next two to three decades. SBIRS, which will replace the 29-year-old DSP system, is designed to support multiple missions, including missile warning and detection, missile defense, technical intelligence, and battlespace characterization. The system includes two major components, SBIRS High and Low (Fig. 3).

SBIRS High, scheduled for initial deployment in 2004, will employ satellites in geosynchronous orbit as well as hosted payloads in highly elliptical orbits. The SBIRS consolidated ground system will be developed in three increments phased to support DSP continental U.S. processing consolidation and the SBIRS High and SBIRS Low constellation deployments. The SBIRS Mission Control Station is located in a new facility located at Buckley Air National Guard Base in Colorado.

The SBIRS Low constellation of low-altitude satellites, which will perform midcourse tracking, is planned for initial deployment in 2006. At this point, SBIRS Low is planned as a primary part of comprehensive missile defense. However, its detailed ground facilities and operations have not yet been set.

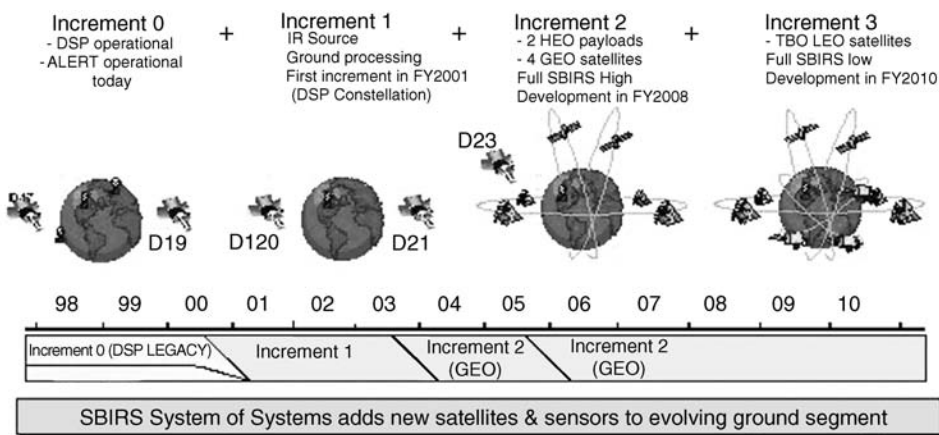


Figure 3. The Space-Based Infrared System (SBIRS) program will provide the nation with critical missile defense and warning capability well into the twenty-first century. Figure courtesy USAF Space and Missile Systems Center. Picture can be found at http://www.losangeles.af.mil/SMC/PA/Fact_Sheets/SBIRS.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Missile warning functional operations—into which DSP and missile warning radars and in the future SBIRS will feed, are the responsibility of the United States Space Command and NORAD. The heart of these operations is in the Cheyenne Mountain Operations Center, Colorado Springs, Colorado. The Center collects data from a worldwide system of satellites, radars, and other sensors and processes that information to support critical NORAD and U.S. Space Command missions. For the NORAD mission, the Cheyenne Mountain Operations Center provides warning of ballistic missile or air attacks against North America, assists the air sovereignty mission for the United States and Canada, and, if necessary, is the focal point for air defense operations to counter enemy bombers or cruise missiles. In support of the U.S. Space Command mission, the Cheyenne Mountain Operations Center provides a day-to-day picture of precisely what is in space and where it is located. The Cheyenne Mountain Operations Center also supports space operations, providing critical information such as collision avoidance data for Space Shuttle flights and troubleshooting satellite interference problems.

Cheyenne Mountain Operations Center (CMOC). Cheyenne Mountain operations are conducted by six centers staffed 24 hours a day, 365 days a year. The centers are the Command Center, the Air Defense Operations Center, the Missile Warning Center, the Space Control Center, the Combined Intelligence Watch Center, and the Systems Center (Fig. 4).

The Command Center is the heart of operations in Cheyenne Mountain. In this center, the Command Director and his crew serve as the NORAD and U.S. Space Command Commander in Chief's direct representative for monitoring, processing, and interpreting missile, space, or air events that could have operational impacts on our forces or capabilities or could be potential threats to North America or U.S. and allied forces overseas. The Command Center is linked directly to the National Command Authorities of both the United States and



Figure 4. The Cheyenne Mountain Combat Operations Center (COC) is protected by a series of massive steel doors embedded in reinforced concrete located in a tunnel more than a thousand feet underground. Figure courtesy Cheyenne Mountain Operations Center. Picture can be found at <http://www.cheyennemountain.af.mil>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Canada as well as to regional command centers. When required, the Command Director must consult directly with the NORAD and U.S. Space Command Commander in Chief for time-critical assessments of missile, air, and space events to ensure that the Commander in Chief's response and direction are properly conveyed and executed.

The CMOC comprises the largest and most complex command and control network in the world. The system uses satellites, microwave radio routes, and fiber optic links to transmit and receive vital communications. Two blast-hardened microwave antennae and two underground coaxial cables transmit the bulk of electronic information. Most of this information is data sent from the worldwide space surveillance and warning network directly to computers inside the Mountain. Redundant and survivable communications hot lines connect the Command Center to the Pentagon, White House, U.S. Strategic Command, Canadian Forces Headquarters in Ottawa, other aerospace defense system command posts, and major military centers around the world.

With respect to space operations and systems, the primary user of space-based warning data is the Missile Warning Center. It uses a worldwide sensor and communications network to provide warning of missile attacks, either long or short range, launched against North America or North American forces overseas. The Missile Warning Center is divided into "strategic" and "theater" sections. The strategic section focuses on information regarding missile launches anywhere on Earth that are detected by the strategic missile warning system and could be a potential threat to Canada or the United States. The theater section focuses on short-range missile launches processed by a Theater Event System that monitors missile launches in areas or theaters that could threaten U.S./allied forces, such as when Iraqi SCUD missiles threatened coalition troops in Operation Desert Storm. Cheyenne Mountain's capabilities to provide timely and accurate warning and cueing for defensive systems such as the Patriot batteries have improved considerably since Desert Storm and continue to improve as new computer and communications systems are added to the Cheyenne Mountain Operations Center (15).

Another key user and driver for space-based data is the Space Control Center. Its detailed functions are to be discussed later. It supports the space control missions of space surveillance and protection of our assets in space. This center was formed in March 1994 through the combination of the Space Surveillance Center and Space Defensive Operations Center. The Space Control Center's primary objective in performing the surveillance mission is to detect, track, identify, and catalog all man-made objects in space. The Center maintains a current computerized catalog of all orbiting space objects, charts objects, charts present positions, plots future orbital paths, and forecasts times and general locations for significant objects reentering Earth's atmosphere. Since 1957, more than 24,000 space objects have been cataloged; many of them have since reentered the atmosphere. Currently, there are about 8000 on-orbit objects being tracked by the Space Control Center. The Center's protection mission is accomplished by compiling information on possible hostile threats that could directly or indirectly threaten U.S./allied space assets. This information is then analyzed to determine the effects/impacts of these threats to our assets in space, so that timely warning and countermeasure recommendations can be made. A good

example of this mission is our constant protection of the Space Shuttle while in orbit, by providing collision avoidance information to NASA (15).

Combined Intelligence Watch Center. There are more than 1100 military and civilian personnel working in the Mountain. Although Cheyenne Mountain would probably not survive a direct hit from today's accurate and high-yield nuclear weapons, it could survive a lower yield nuclear and conventional weapons impact. It is also well protected against other actions such as sabotage and terrorism. It is self-sustaining, capable of providing its own power, water, air, and food for up to 800 people for 30 days.

Space Force Enhancement: Meteorological Satellites. The Defense Meteorological Satellite Program (DMSP) was originally the DOD weather observation system. It is currently composed of two satellites placed in Sun-synchronous orbits 450 nautical miles above Earth. The Air Force's Space Systems Division began developing weather satellites and associated ground stations and terminals during the 1960s (16). The existence of a military weather system was classified secret until 1973. Since that time, all elements of the system have been upgraded several times: the current system is the military's sixth generation. As for many other initially classified systems, DMSP developed its ground and space operations system as a "stovepipe," largely separate from other satellite systems (Fig. 5).

Currently, DMSP Block 5D-2 satellites provide cloud imaging over Earth's surface during both day and night. The satellite also carries a number of additional weather and space environment ("space weather") sensors. The visible and infrared sensors collect images of global cloud distribution across a 3000-km swath during both daytime and nighttime conditions. The microwave imager and sounders (SSMIS) are capable of recording ocean wind speed, ice age and density,



Figure 5. The Defense Meteorological Satellite Program (DMSP) has been collecting weather data for U.S. military operations for almost four decades. At all times, two operational DMSP satellites are in polar orbit at about 458 nautical miles (nominal). The primary weather sensor on DMSP is the Operational Linescan System which provides continuous visual and infrared imagery of cloud cover over an area 1600 nautical miles wide. Figure courtesy USAF Space Command. Picture can be found at <http://www.peterson.af.mil/hqafspc/Library/FactSheets/FactSheets.asp?FactChoice=4>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

water content in soils, and other scientifically useful data. The SSMIS covers only half the swath width of the infrared and visible sensors. This coverage limitation results in full coverage of Earth's polar regions above 60°, twice daily, and the equatorial region, daily. The space environmental sensors record along-track plasma densities, velocities, composition, and drifts.

The Block 5-D system was originally operated as part of the Air Force's 50th Space Wing as the 6th Space Operations Squadron at Offutt Air Force Base, Nebraska; satellite control stations are situated at Loring Air Force Base, Maine, and Fairchild AFB, Washington. DMSP could also be operated through the AFSCN for early-orbit checkout and satellite emergencies. However, as part of civilian and military meteorological satellite consolidation begun in 1994, DMSP is now operated through the National Oceanic and Atmospheric Administration's (NOAA) Suitland, Maryland, operations center, and the Loring, Maine, uplink-downlink site is closed.

The DMSP spacecraft operates in two modes. It can directly downlink data to tactical terminals (Mark IV Series Transportable Terminals, AN/SMQ-11 Shipboard Receiving Terminals, and Rapid Deployment Information Tactical Terminals—RDITs). The Tactical Terminals receive DMSP mission data in real time. Alternatively, data may be stored on board the satellite on tape recorders for "store and forward" downlink primarily to the Fairchild ground station and retransmission to users including Air Force Global Weather Central (AFGWC) and the Navy Fleet Numerical Meteorological Oceanography Center (FNMOC).

AFGWC, located at Offutt Air Force Base in Omaha, Nebraska, is the primary strategic user and distributor of DMSP satellite data destined for the Air Force and Army. FNMOC, located at Monterey, California, distributes DMSP data to the Navy and Marine Corps. Although neither AFGWC nor FNMOC are DMSP agencies, both receive and process DMSP data in combination with meteorological, solar-geophysical, and oceanographic observations from other sources. They disseminate such environmental information in various forms to the DOD and other governmental agencies, as required.

A 1994 U.S. White House decision mandated convergence of DMSP with NOAA's Polar-Orbiting Operational Environmental Satellite (POES) program. The goal of the converged program is to reduce the cost of acquiring and operating polar orbiting operational environmental satellites, while continuing to satisfy U.S. military, civil, and national security requirements. As part of this goal, the converged program will incorporate appropriate aspects of both DMSP and NASA's Earth Observing System. The converged system on-orbit architecture will consist of three low Earth orbiting satellites versus the current four satellites (two civilian and two military). The orbits of the three satellites will be evenly spaced throughout the day to provide timely data refresh. The nominal equatorial crossing times of the satellites will be 5:30, 9:30, and 1:30 (17).

Space Force Enhancement: Navigation. The U.S. Navigation mission is executed by the Global Positioning System (GPS). This program, due to its status as an essential "global utility," is arguably the military's most significant space program. GPS grew out of two predecessors—an Air Force technology program started in the late 1960s, 621B, and a parallel U.S. Naval Research Laboratory program called Timation (16). 621B proposed a constellation of about 20 inclined geosynchronous satellites, and Timation suggested a constellation of 21 to 27

satellites in a medium altitude Earth orbit (MEO). Elements of both programs were combined in 1973 to create the initial GPS concept. This initial concept would employ the signal structure and frequencies of 621B and the MEO orbits of the Timation proposal (Fig. 6).

The GPS program began in 1973, and it was acquired in three phases—validation, development, and production. Ironically, the USAF was a reluctant program manager and had to be repeatedly forced to proceed with system development. Block I navigation satellites and a prototype control segment were built and deployed, and advanced development models of various types of user equipment were built and tested. During the development phase, additional Block I satellites were launched to maintain the initial satellite constellation, and Block II satellites were tested and began deployment. In addition, an operational control segment was activated under the Air Force Space Command's 50th Space Wing at Falcon (now Schriever) Air Force Base, Colorado (18). During the production phase, a full constellation of 24 Block II and IIA satellites was deployed, and user equipment was produced and put into the field. The full constellation of Block II and IIA satellites completed in March 1994 allowed the system to attain full operational capability in April 1995.

The CONTROL segment is comprised of five monitor stations located in Hawaii, Kwajalein, Ascension Island, Diego Garcia, and Colorado Springs, three ground antennae located at Ascension Island, Diego Garcia, and Kwajalein), and a master control station (MCS) located at Schriever AFB in Colorado. All monitor stations except Hawaii and Colorado (Schriever) are equipped with ground antennae for communications with the GPS satellites. The monitor stations passively track all GPS satellites in view and collect ranging data from each satellite. This information is passed on to the MCS where the satellite orbital



Figure 6. The Navstar Global Positioning System (GPS) is a constellation of orbiting satellites that provides navigational data to military and civilian users all over the world. The system is operated and controlled by members of the 50th Space Wing, located at Schriever AFB, Colorado. Figure courtesy USAF Space Command. Picture can be found at <http://www.peterson.af.mil/hqafspc/Library/FactSheets/FactSheets.asp?FactChoice=9>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

parameter data (ephemeris) and clock parameters are estimated and predicted. The MCS periodically uploads the updated ephemeris and clock data to each satellite for retransmission in the navigation. The USER segment consists of antennae and receiver-processors that provide positioning, velocity, and precise timing to the user.

The GPS system's importance to global commercial and civil uses was recognized during the 1990s. GPS provides precise positions for military, civil, and commercial purposes worldwide, but it may be even more important as the "global clock." As a means of "time-transfer," users everywhere rely on GPS to time everything precisely, from communications circuits to bank transactions, to within a few billionths of a second. To see how important this is, consider what happened when a real error occurred in 1996. A satellite controller at the Air Force's GPS control center accidentally put a bad time into just one of GPS's 24 satellites. The bad time was broadcast for only six seconds before automatic systems detected it and shut the satellite signal down. Nonetheless, more than 100 of the 800-plus cellular telephone networks on the U.S. East Coast—which rely on precise GPS-provided timing to work—failed. Some took hours and even days to recover. GPS directly produces several tens of billions of dollar revenue for the United States yearly—indirectly it produces many times this amount.

In recognition of the criticality of GPS as a global utility, the U.S. government has implemented a number of important upgrades and initiatives following a March 1996 White House decision. In 1999, the United States decided to remove the deliberate signal degradation ("Selective Availability"), which limited nongovernmental users to considerably worse signal performance. In addition, the U.S. Department of Defense has officially taken on modernization of the GPS system to provide additional precise signals for civil and commercial users in the second "military" frequency and to provide a third civil frequency. Moreover, military users will now have a separate, more protected signal structure within the two original GPS bands. These upgrades are to be implemented on 12 of the current block of GPS IIR satellites and future IIF. Finally, a new GPS system development, GPS III will launch late in the decade to provide greatly improved signal structure and strength. In recognition of the importance of the timing function, the U.S. Naval Observatory has placed an alternate master clock time standard at Schriever Air Force Base, and the GPS program is building a protected alternate master control station at Vandenberg Air Force Base, California.

Space Force Enhancement: Communications. Communications support is one of the most long-standing and critical systems supporting war fighters from and through space. Despite the rapid growth in fiber-based communications, military support is usually required in locations where ground-based fiber links are unavailable or have been cut. Thus, military satellite communications systems (MILSATCOM) remain the heart of national security communications networks.

Before delving into individual system operating concepts, it is necessary to understand the overarching DOD approach to communications. The top-level architecture is contained in the U.S. Department of Defense Information Systems Network (DISN), which is the responsibility of the Defense Information Systems Agency (DISA), part of the Office of the Secretary of Defense. DISN prescribes a global network integrating Defense Communications System assets,

MILSATCOM, Commercial SATCOM initiatives (which now include the troubled IRIDIUM system), leased telecommunications systems, dedicated DOD Service and Defense Agency networks, and mobile/deployed networks, such as the consolidated worldwide enterprise level telecommunications infrastructure, for example, that provides the end-to-end information transfer component of what's known as the "Defense Information Infrastructure" (DII) (19). The DISN provides rapid information access "to allow any warrior to perform any mission, anytime, any place in the world, based on information needs" (19). DISA provides top-level policy and architecture support and is the primary route for providing commercial communications support to war fighters. Conversely, MILSATCOM operation and support, subject to DISA policy guidance, is provided through the United States Space Command and its components.

Various satellite communications systems have been developed. The first of these, the Initial Communications Satellite Program (IDCSP), began in 1962. It consisted of small, 100-pound satellites launched in clusters. A total of 26 such satellites were placed in orbit between June 1966 and June 1968 (16). It was superseded by the more sophisticated Defense Satellite Communications System, Phase II (DSCS II). The first two DSCS II satellites were put into GEO orbit in November 1973, and a total of 16 were built and launched during the life of the program; the final launch was in 1989. DSCS II and its successor DSCS III use super-high-frequency (SHF) transponders.

The U.S. Air Force began launching DSCS III satellites in 1982. DSCS III has more flexible coverage than DSCS II and provides increased power, which can be tailored to suit the needs of different site user terminals. There is an expected economic payoff from the operation of the DSCS III satellites; they are expected to have a useful lifetime twice that of the DSCS II satellites. DSCS III can also be used to disseminate emergency action and force direction messages to nuclear capable forces. The system can provide worldwide secure voice and high data rate communications. The system is also designed to resist jamming (20).

As in many other military systems, DSCS III separates satellite system (bus) operations from payload operations. Air Force Space Command units operate the DSCS bus, the 50th Space Wing's 3rd Space Operations Squadron at Schriever Air Force Base, Colorado, and the 5th Space Operations Squadron at Onizuka Air Force Base, California. Conversely, the payloads are operated by Army Space Command crews at the major communications uplink and downlink sites in the United States and overseas. Allocation of communications channels shifted in 1998 from the Joint Staff in the Pentagon to the United States Space Command's Space Operations Center (SPOC) at Peterson AFB, Colorado.

The U.S. Navy has also invested heavily in space-based communications systems. During the 1960s, the military saw the need to transmit data to much smaller terminals than those required by the DSCS VHF system. In February 1969, Lincoln Laboratory's Tactical Communications Satellite (TACSAT I) was launched and proved the feasibility of communications with small tactical terminals in the UHF band. This paved the way for the Navy's Fleet Satellite Communications System (FLTSATCOM). The U.S. Navy conducted overall management, and the Air Force managed the system acquisition. Between February 1978 and September 1989, eight FLTSATCOMs were built and launched. However, only six became fully operational in orbit. The Navy supplemented its

FLTSATCOM system during this period and the 1990s with leased commercial UHF satellite services (16).

The Navy began replacing its UHF satellite constellation during the early 1990s with a constellation of customized Hughes 601 spacecraft known as the UHF—Follow-On (UFO) satellites (Fig. 7). So, yes, the U.S. DOD does have a not-so-secret fleet of UFOs! These satellites were purchased in a novel “services in orbit” contract that needed minimal government oversight. Nine UFO satellites were successfully launched from 1993–1999.

In March 1996, the U.S. Navy ordered a high-speed, high-power global broadcast system (GBS) to be added to its future UFO satellites. This GBS package provides the military equivalent of “direct broadcast TV.” It is revolutionizing the dissemination of high-capacity data ranging from intelligence imagery to “quality-of-life” television for forward-deployed forces (21).

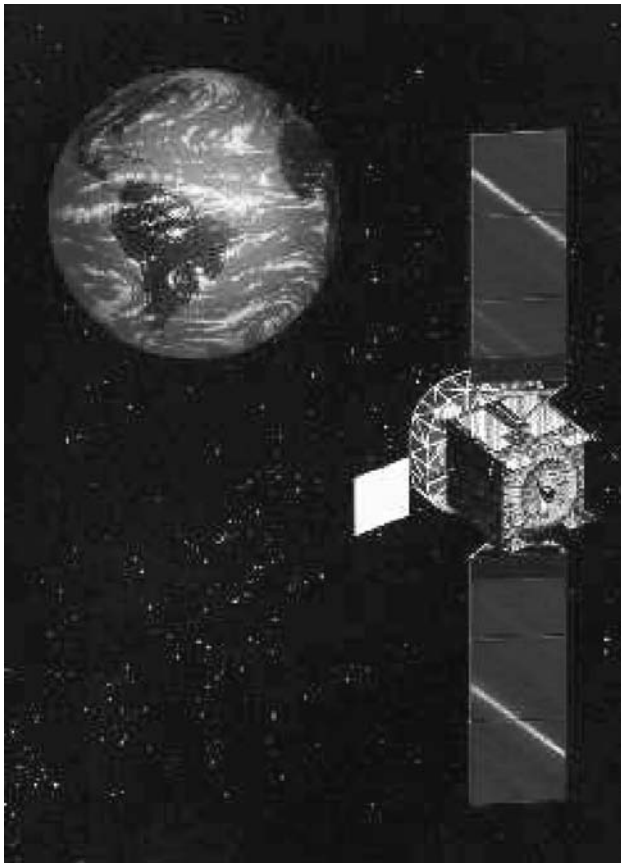


Figure 7. The first UHF F/O was launched 25 March 1993. The Atlas II rocket booster malfunctioned, placing the spacecraft in a dangerously low orbit after efforts by the 3rd Space Operations Squadron, Schriever AFB, Colorado. Each UFO satellite will possess 39 UHF communications channels (a 70% increase over Fleet Satellites). Figure courtesy USAF Space and Missile Systems Center. Picture can be found at <http://www.au.af.mil/au/awc/awcgate/usspc-fs/space.htm>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The FLTSATCOM systems, as well as the UFO satellites, were originally operated by the Air Force Space Command in concert with other military communications systems. However, in 1999, the eight on-orbit UFO satellites were transferred from the Air Force to the Naval Space Command's Naval Satellite Operations Center (NAVSOC) in Dahlgren, Virginia. Also in 1999, the Chief of Naval Operations transferred control of the Navy's commercial satellite services from its research and development command, the Space and Naval Warfare Systems Command, to the Naval Space Command (22).

NAVSOC provides "one-stop shopping" for operational satellite intelligence and communications to deployed U.S. Navy and Marine forces worldwide. In addition, the Navy maintains its own space surveillance capability and sensors as part of the Naval Space Command and its support to the United States Space Command.

To provide additional tactical communications support, the U.S. Department of Defense agreed in late 1999 to continue commercial operations of the troubled Iridium, low Earth orbit network of 70-plus linked communications satellites. In this sense, the U.S. military became an anchor-tenant for the global handheld communications system, which was unable to compete successfully with the burgeoning fiber-optic and cellular services available worldwide. However, the U.S. military's unique requirements for communications in remote, particularly Arctic areas, make this a cost-effective approach to maintaining secure connectivity to deployed forces.

In addition to the long-haul DSCS and commercial satellite users and tactical users of FLTSATCOM and the UFO, there is a third group of users—nuclear-capable forces. These users require extremely high reliability and only very low data rates, as well as very high survivability. Initially, the Air Force provided this capability through its Air Force Satellite Communications System (AFSATCOM). It is carried as an additional payload on FLTSATCOM and other DOD spacecraft. It became operational with both user terminals and satellites in the late 1970s.

Replacement of AFSATCOM began in 1994 with the successful launch and checkout of the first MILSTAR I satellite. MILSTAR is a worldwide, survivable, highly jam-resistant communications satellite system that enables the U.S. President and his Secretary of Defense to communicate with tactical and strategic nuclear forces (Fig. 8). MILSTAR II was initiated in 1993; it will provide protected, medium data rate, secure communications capabilities. To date a total of five spacecraft have been launched and are operational. The last of them was launched on 15 January 2002. A launch of the seventh MILSTAR satellite is scheduled for 4 November 2002. If successful, this will make a constellation with a total of six MILSTAR spacecraft in orbit (23).

MILSTAR is another program operated independently of other satellite systems. It includes both tactical terminals and a central dedicated control capability at Schriever AFB, Colorado, under the 50th Space Wing's 4th Space Operations Squadron. A reserve satellite operations squadron at Vandenberg AFB, California, provides backup master control.

Space-Based Reconnaissance. Intelligence uses of space are the responsibility of the National Reconnaissance Office (NRO) as discussed before. The details of these satellite's operations and facilities remain classified. However,

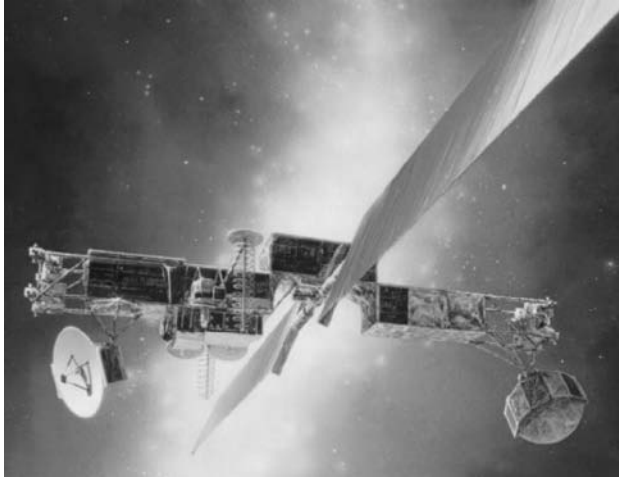


Figure 8. Milstar is a joint service satellite communications system that provides secure, jam-resistant, worldwide communications to meet essential wartime requirements for high priority military users. The multisatellite constellation will link command authorities with a wide variety of resources, including ships, submarines, aircraft, and ground stations. Figure courtesy USAF Space Command. Picture can be found at <http://www.peter-son.af.mil/hqafspc/Library/FactSheets/FactSheets.asp?FactChoice=14>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

most of them use dedicated, system-specific command and control and data distribution systems and sites.

Space Operations. The second major military space function includes all of those efforts and facilities needed to launch military systems into space, control them while in orbit, deorbit or decommission them when their missions are complete, and distribute mission data. For this discussion, we will focus on the second part of these activities—operating satellites in orbit. As discussed previously, U.S. military satellite operations are a combination of system-specific dedicated command and control systems and use of the DOD-wide common satellite control network known as the Air Force Satellite Control Network (AFSCN).

The AFSCN tracks DOD satellites, receives and processes telemetry and in some cases data transmitted by them, and sends commands to them. At its peak in the mid-1990s, the AFSCN consisted of two control nodes, two scheduling facilities (one at each node), nine remote tracking sites, and communications links connecting them. As we enter a new century, the scheduling and control node at Onizuka AFB, California, is being phased out, and all activities are to be concentrated at Schriever AFB, Colorado, under the command of the Air Force Space Command's 50th Space Wing. One of the remote tracking sites, at the Seychelles, Indian Ocean, was also shut down in the late 1990s.

The common user element of the AFSCN was originally activated to support the Discoverer program of the late 1950s and early 1960s. An interim satellite control facility was initially established in Palo Alto, California, in January 1959.

By June 1960, a permanent control center had been established in Sunnyvale, California. It was originally named the Satellite Test Annex. Since then, it has been renamed numerous times, and finally, it is called the Onizuka Air Force base in honor of Air Force astronaut Ellison Onizuka killed in the 1986 Challenger Space Shuttle explosion. It was generally known as the “Blue Cube” because it was very visible from nearby freeways as a large, blue-painted, windowless building. Tracking facilities were established at nine different locations between 1959 and 1961 to complement the Sunnyvale control center (16).

Due in large part to the Sunnyvale facility’s location near major public thoroughfares and its unfortunate placement near three active fault lines, concern grew during the 1970s about its vulnerability as a single node failure point. Thus, a second control center was added—the Consolidated Space Operations Center (CSOC) located in what is now known as Schriever AFB. The Secretary of Defense authorized CSOC in 1979. Originally, it was to consist of two parts—a satellite operations complex (SOC) for on-orbit satellite control and a Shuttle Operations Planning Center (SOPC) for planning and controlling DOD Space Shuttle missions. However, after the Challenger accident and cancellation of DOD Space Shuttle use, the SOPC was itself canceled in 1987. The CSOC began operations in 1989 and was turned over to the Air Force Space Command after initial operational test and evaluation in 1993. The Onizuka site is in the process of being phased out, and most of its functions are now transferred to an enlarged facility at Schriever AFB. In the late 1990s, the U.S. Navy’s independent satellite control facilities were merged with the AFSCN, which now serves as the DOD’s sole common user satellite control network.

The AFSCN remains an impressive capability. However, despite continued upgrades for improved automation and reliability, it remains a manpower-intensive system requiring more than 1000 contractor and government people to maintain and operate it.

The biggest challenge facing the AFSCN today is in satellite control frequencies. The AFSCN uses a U.S. military-unique frequency structure called the Space–Ground Link System or SGLS. This frequency structure was designed in the 1950s to provide a robust, jam-resistant uplink and downlink capability. It uses 20 separate channels for uplink between 1.75 and 1.85 GHz. Similarly, SGLS uses 20 downlink channels around 2.3 GHz. However, as pressure grows on the overall frequency spectrum and because the U.S. military does not have global assignment of these frequencies, it is faced with an increasingly urgent necessity to transition to other satellite control frequencies or approaches. For the long-term, space-to-space options similar to those used in NASA’s Tracking and Data Relay Satellite System (TDRSS) appear most desirable. For the interim, the U.S. military is considering transiting to the Unified S-band system used by NASA.

Space Control and Satellite Tracking. Space control is the third major U.S. military mission area. It consists of four elements: (1) space surveillance to provide critical space situational awareness; (2) protection methods to ensure that U.S., allied and commercial systems vital to U.S. national security operations are not interfered with; (3) preventive means to ensure that adversaries do not use U.S. systems such as the GPS to assist in their military operations, and finally; (4) negation means to deny adversaries use of their own space systems for

aggression. Currently, almost all U.S. work focuses on the surveillance and protection aspects of space control.

The United States Space Command holds overall responsibility for space control operations. Central to this function is the Space Control Center (SCC) in Cheyenne Mountain, Colorado. Its primary function is to maintain tracking data on all objects in orbit and to determine if threats exist to national security space operations. The SCC is manned jointly by U.S. Space Command personnel who maintain overall responsibility for identifying threats to U.S. and allied space systems and Air Force Space Command's 21st Space Wing 1st Command and Control Squadron that provides tasking to the Space Surveillance Network's sensors. The U.S. Navy Space Command operates the Alternate Space Control Center in Dahlgren, Virginia.

Space Control and Satellite Tracking—The Space Surveillance Network.

Space surveillance involves detecting, tracking, cataloging, and identifying man-made objects orbiting Earth. These objects include active/inactive satellites, spent rocket bodies, or fragmentation debris. Space surveillance accomplishes the following:

- predicts when and where a decaying space object will reenter Earth's atmosphere;
- prevents a returning space object, which to radar looks like a missile, from triggering a false alarm in missile-attack warning sensors of the U.S. and other countries;
- charts the present position of space objects and plots their anticipated orbital paths;
- detects new man-made objects in space;
- produces a running catalog of man-made space objects;
- determines which country owns a reentering space object;
- informs NASA whether or not objects may interfere with the Space Shuttle and International Space Station.

The command accomplishes these tasks through its Space Surveillance Network (SSN) of U.S. Army, Navy, and Air Force operated ground-based radars and optical sensors at 25 sites worldwide (Fig. 9).

The SSN has been tracking space objects since 1957 when the Soviets opened the Space Age by launching Sputnik I. Since then, the SSN has tracked more than 24,500 space objects orbiting Earth; two thirds of these have reentered Earth's atmosphere and disintegrated or impacted Earth (24).

SSN Sensors. Due to network capacity limitations (number of sensors, location, availability), the SSN uses a "predictive" technique to monitor space objects, that is, it spot-checks them rather than tracking them continually. Following is a brief description of each type of sensor.

Phased-array radars can maintain tracks on multiple satellites simultaneously and scan large areas of space in a fraction of a second. These radars have no moving mechanical parts to limit the speed of the radar scan—the radar energy is steered electronically (Fig. 10).

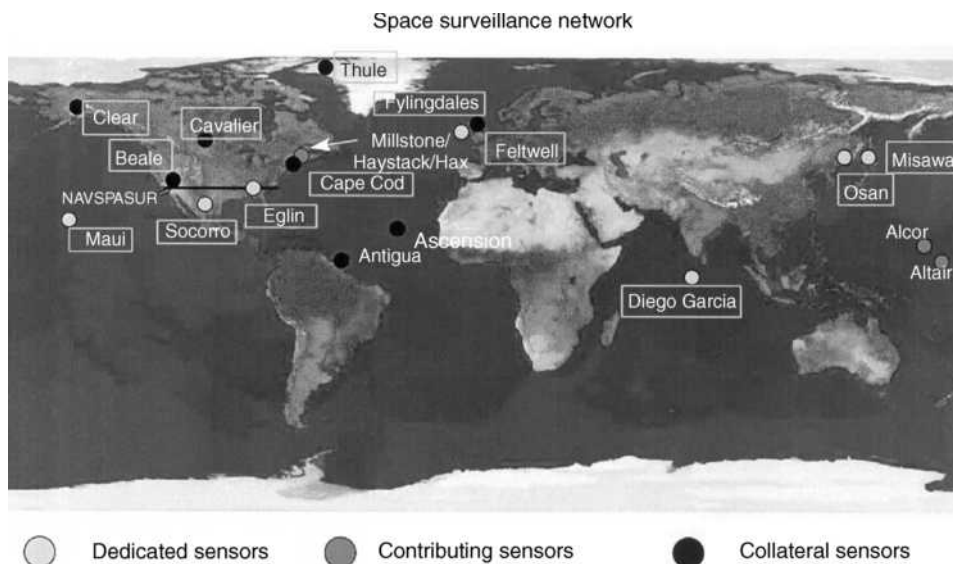


Figure 9. The Space Surveillance Network is a worldwide network of ground- and space-based sensors that has radar and ground stations located on every continent. Figure courtesy USAF Air University. Picture can be found at <http://www.au.af.mil/au/awc/awcgate/usspc-fs/space.htm>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Conventional radars use immobile detection and tracking antennae. The detection antenna transmits radar energy into space in the shape of a large fan. When a satellite intersects the fan, the energy is reflected back to the antenna, triggering the tracking antenna. The tracking antenna, then, locks its narrow beam of energy on the target and follows it to establish orbital data (Fig. 11).

The Ground-Based Electro-Optical Deep Space Surveillance System (GEODSS) consists of three telescope sensors linked to a video camera (Fig. 12). The video cameras feed their space pictures into a nearby computer, which drives a display scope. The image is transposed into electrical impulses and recorded on magnetic tape. This is the same process used by video cameras. Thus, the image can be recorded and analyzed in real time (24).

In 1998, the Air Force Space Command began using the missile defense experimental spacecraft "MSX" to explore the use of space-based space surveillance sensors. This has been a very successful experiment in detecting and tracking geosynchronous (GEO) orbiting satellites. Since its initial use, the MSX sensor has succeeded in lowering the number of "lost" satellites in GEO by more than a factor of 2.

Combined, these sensors make up to 80,000 satellite observations each day. This enormous amount of data comes from SSN sites in Maui, Hawaii; Eglin, Florida; Thule, Greenland; and Diego Garcia, Indian Ocean. The data is transmitted directly to the SCC via satellite, ground wire, microwave, and telephone. All available methods of communication are used to ensure that a backup is readily available, if necessary.



Figure 10. The Space Surveillance Network is composed in part of ground-based radars, like this precision acquisition vehicle entry and phased array warning system (PAVE PAWS), which detects and tracks sea-launched and intercontinental ballistic missiles. Their secondary mission is to track objects in space. Figure courtesy USAF Space Command. Picture can be found at <http://www.peterson.af.mil/hqafspc/library/FactSheets/FactSheets.asp?FactChoice=15>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Space Control Center (SCC). The SCC in Cheyenne Mountain Air Station is the terminus for the SSN's abundant flow of information. The SCC houses large, powerful computers to process SSN information and accomplish the space surveillance and space control missions.

The NAVSPACCOM provides the site and personnel for the Alternate SCC (ASCC). The ASCC would take over all operations in the event the SCC could not function. This backup capability is exercised frequently.

Orbital Space Debris. USSPACECOM tracks about 8000 man-made space objects, baseball-size and larger, that orbit Earth. The space objects consist of active/inactive satellites, spent rocket bodies, or fragmentation. About 7% are operational satellites, 15% are rocket bodies, and about 78% are fragmented and inactive satellites; the rest are debris. USSPACECOM is primarily interested in the active satellites but also tracks space debris. The SSN tracks space objects, which are as small as 10 cm in diameter (baseball size) or larger.



Figure 11. PIKE, a Remote Tracking Station at Schriever AFB, Colorado, looks like a giant golf ball to casual observers. PIKE is operated by the 22nd Space Operations Squadron. Figure courtesy USAF Space Command. Picture can be found at http://www.peterson.af.mil/hqafspc/Library/almanac/pg12/almanac_12.htm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



Figure 12. GEODSS sites play a vital role in tracking deep space objects. More than 2500 objects, including geostationary communications satellites, are in deep space orbits more than 3000 miles from Earth. Three operational GEODSS sites report to the 18th Space Surveillance Squadron, Edwards AFB, California; Socorro, Natingham; Maui, Hawaii; and Diego Garcia, British Indian Ocean Territories. Figure courtesy USAF Space Command. Picture can be found at <http://www.peterson.af.mil/hqafspc/Library/FactSheets/FactSheets.asp?FactChoice=8>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Although 8000 space objects seems like a large number, in the 800-km orbital belt, there are normally only three or four items in an area roughly equivalent to the airspace over the continental U.S. up to an altitude of 30,000 feet. Therefore, the probability of collision between objects is very small (24).

During the 23 March 2001 reentry of the Russian MIR space station, the SSN proved its worth as the only worldwide sensor network capable of monitoring the MIR's precise location and configuration nearly continuously.

Space Control—Protection, Denial, and Negation. Although space surveillance is the basis of all space control functions, the uniquely military functions of protecting U.S. and other friendly satellites from hostile interference, denial of use of friendly satellite systems such as GPS, and negation of hostile space capabilities are uniquely military functions. During the 1990s, there have been a number of hostile interferences with communications satellites. The best known was the apparent deliberate jamming of a Tongan-leased Chinese communications satellite by Indonesia. The latter had claimed the same GEO operating slot and frequency usage. The U.S. Department of Defense focused in 1999 on a broad area review of space control and concluded that effective space surveillance was key but that far more attention was also needed in the areas of protection and denial.

To date, little is possible in the realm of satellite protection. The Space Control Center attempts to fuse information on potential satellite attacks based on reports from satellite operators, intelligence information, and satellite users. However, this data is often not timely enough to provide effective protection of satellite systems. A related satellite protection initiative is to increase force protection of satellite ground facilities and communications links. As for most military systems in the early twenty-first century, these facilities are increasingly vulnerable to terrorist attack. As military reliance on these space systems grows, their physical protection will take on new urgency.

Related to protection of space capabilities is the ability to deny an adversary use of friendly military and commercial systems. Key among these capabilities is the Global Positioning System. As previously outlined, the system is capable of encrypting its high accuracy data—that which could be used by a hostile group for precise guidance of a munition—and only providing degraded publicly available data. In 1998, the U.S. government decided to remove this degradation. However, it reserves the right to reactivate the degradation in the event of hostilities. Currently, the GPS degradation must be implemented globally at the direction of the satellite control element at Schriever Air Force Base, Colorado. However, the Department of Defense has been directed to develop means to limit degradation to local crisis areas and is actively pursuing a variety of means to do so.

Another denial capability is to prohibit hostile nations or groups from receiving commercial or other space-based imagery, which could be of military use during crisis or conflict. Currently, this is accomplished through “shutter-control” agreements with commercial and foreign entities licensed to use U.S. produced systems or components—which is most of those currently available. However, as more and diverse groups develop space-based imaging and other surveillance capabilities, it will become impossible to control access to high quality imagery and other surveillance products. This will drive governments to develop more active means of denial.

Negation of on-orbit space systems remains a politically sensitive topic. In the 1970s, the Soviet Union developed an antisatellite weapon (ASAT). In the early 1980s, the United States tested its own antisatellite system—an air-launched miniature homing vehicle (MV-ASAT) which would crash into the target satellite. During its brief life as a prototype operational system, command and control for the system was done through the SCC's predecessor, the Space Defense Operations Center (SPADOC) within Cheyenne Mountain. Both the Soviet and U.S. operational ASATs have been discontinued. However, the United States has publicly stated that it is pursuing negation capabilities and reserves the right to deploy them should the military and world situation so warrant. As for the MV-ASAT, command and control for such future systems would devolve through the U.S. Space Command's Space Control Center through the service component actually operating the weapon system.

One of the most likely near-term negation systems is the ground-based laser. Both the United States and the former Soviet Union have developed ground-based lasers capable of damaging LEO satellites. The Mid-Infra Red Advanced Chemical Laser (MIRACL) at the White Sands Missile Range—operated by the U.S. Army Space and Missile Defense Command, the U.S. Army service component of the U.S. Space Command—retains a contingency ASAT capability (25). Other nations such as China are believed to be actively pursuing ground-based laser ASAT systems.

Future Directions

As we enter the twenty-first century, there is considerable uncertainty about the direction of future military use of space, both in the United States and throughout the world. Today, the United States stands alone in having a substantial military space organization and infrastructure. However, Russia has retained much of its former Soviet space potential and has shown signs of revitalizing its military space organization and capabilities. Several countries within the European Community—particularly France—have a close relationship between their civil space effort and a growing military use of space. China and India both have growing space access and use capabilities and have devoted increasing portions of these efforts toward national security missions. Smaller nations, such as Israel and even Chile, have developed and launched military space systems.

It is difficult to predict the future direction of military space, but several trends are apparent. In the area of space operations, there are two emerging trends. First is the growing need to maintain continuous contact with space systems. For GEO systems, this is relatively straightforward because a ground antenna at the subsatellite point can maintain a 24-hour a day lock on the satellite. However, these links are vulnerable to both ground attack and electronic interference (jamming). The problem is much more difficult for LEO and MEO satellites which do not stay in continuous view of a ground station and rely on a network of ground sites situated around the world. Correspondingly, they are in contact only a few hours a day—usually for only a few minutes at a time as they rise and fall from the field of view of each station. Thus, considerable effort is focused on developing continuous satellite contact through space-to-space

communications links. Such capabilities have long been embodied in NASA's TDRSS. However, future military space-based satellite control systems must be cognizant of the potential for hostile interference. For this reason, the military is looking at two approaches. The first relies on using radio frequencies, which cannot be propagated effectively through the atmosphere—thus, the satellite-to-satellite links are not susceptible to ground-based interference. Alternately, space-to-space communications may be via very narrowband laser cross-links—making it very difficult for an adversary to enter or jam the communications link.

Another potential trend in space operations is the likely emergence of fully reusable space access systems. The U.S. DOD refers to such systems as “spaceplanes.” Affordability is a concern, but fully reusable space access systems would allow aircraft-sortie-like access to space. If this capability were realized, key military space systems might no longer be permanently stationed in space but would be put in place during a crisis or conflict and potentially recovered after their mission was completed. For permanently stationed space systems such as GPS, reusable sortie access to space would allow these systems to be much more easily repaired or upgraded or replaced in orbit. Today, it will take more than 10 years to replace the existing GPS system fully with a system currently being built to add new military and civilian frequency service. With reusable space access and reconfigurable space systems, this change might be accomplished in a few months versus a decade or more. The emergence of the reusable spaceplane would thus offer a complete revolution in space operations. This revolution may not be far off, but it will take renewed investment in “X-vehicles” by NASA that has lately been quick to eliminate such high-risk programs.

Space Control may also be revolutionized by two new developments. The first is the emergence of “microsatellite” technologies. In recent years, a number of groups around the world began constructing very sophisticated space systems weighing 100 kilograms or less. These so-called “microsatellites” can perform many of the scientific functions formerly requiring satellites weighing thousands of kilograms. Microsatellites not only weigh less, they cost less, too. Typical system development costs are in the few millions of dollars versus hundreds of millions for a comparable conventionally sized satellite. Of course a cheap satellite is of less appeal if it still costs a large fraction of \$100 million to launch, as is the case currently. Here, too, a revolution is occurring. Although fully reusable launch vehicles are a decade or more away, microsatellites have been able to take advantage of “free” space on larger launch vehicles. This has been particularly true of the European Ariane launcher—which since the late 1980s has had available space for a number of microsatellites weighing up to about 50 kilograms. The newer Ariane V can launch up to eight secondary microsatellites into GEO transfer orbits on each launch. Typical costs for putting a microsatellite into orbit as a secondary payload are about \$1 million. Thus, not only are low-cost microsatellites feasible, they are “cheap” to launch into LEO orbits and higher orbits as well.

Currently, the most advanced microsatellite development group is the UK's University of Surrey Space Centre. This group has constructed and launched more than 20 small and microsatellite class payloads—most as secondary payloads. Particularly interesting have been the successful efforts of the Surrey concern to transfer microsatellite capability to nations not normally associated

with space programs such as Chile, Thailand, Malaysia, and Korea. Recently, the Surrey Space Centre has developed the next iteration of smaller, equally capable satellites—testing their SNAP-1 “nanosatellite” weighing about 5 kilograms. Although this system costs a few hundred thousand dollars to construct, it is no toy. It is three-axis stabilized and contains a propulsion system and a camera, which is used to take very good images of the primary and secondary payloads alongside of which it was launched.

The microsatellite and nanosatellite revolution is interesting from a scientific perspective, but it is even more significant in its implications for military use of space and space control. It places the ability to access space, inspect objects in space, and even interfere with other objects within the reach of most nations. This greatly complicates all aspects of space control—particularly space surveillance and space system protection. Because the current space surveillance system, of which the U.S. capability is most advanced, is a “tracking” system which looks for a satellite where it’s supposed to be, it is not well suited to detect a satellite that operates in a non-Keplerian fashion. A system that searches large areas of space is needed. Furthermore, particularly the “nanosatellites” are very near the limit of detectability—particularly at higher orbits—for today’s surveillance systems. Thus, the challenge for space control systems of the future is to detect and keep tabs on a large number of very small objects, which might be maneuvering frequently. The solution to this problem appears to be space-based optical and infrared surveillance systems.

Currently, only the Air Force’s MSX satellite conducts space-based surveillance operations. It operates only in a tasked track mode, although it can, in principle, search the entire GEO belt. However, it is a research and development satellite that is expected to fail in the near future. Microsatellite technology does, however, offer a means to put in place a low-cost, all-sky, comprehensive search. The Canadian microsatellite-based Near-Earth Surveillance System (NESS) proposes a 50-kilogram surveillance satellite capable of searching large parts of the sky down to a limiting magnitude of about 19. This would detect a nanosatellite out to GEO altitudes.

The emergence of many nations that have microsatellite and nanosatellite technology will change greatly the current approach to space surveillance. Conversely, the same microsatellite technology offers near real-time, all-sky, space surveillance.

The emergence of easy-to-maneuver microsatellites further presses the case for space-to-space continuous communications links. For LEO satellites, as discussed before, this may take the form of a TDRSS-type system. However, continuous space contact might be maintained with a LEO satellite through one of the new multisatellite communications systems such as the LEO Iridium constellation.

Space Force Application. No nation currently uses space to apply military force—other than the significant exception of long-range ballistic missiles—but this situation may soon change. The United States is developing various concepts for space-based missile defense systems. As the offensive missile defense threat grows, as many believe it will, pressure for a global missile defense versus simply a national or theater missile defense will also grow. These global missile defenses might be based on a constellation of space-based laser platforms or could consist

of dozens to thousands of small interceptor missiles in orbit. Conversely, some advocate replacing nuclear-armed ballistic missiles with precision nonnuclear weapons launched on need into or through space to their targets. The United States Space Command was merged with United States Strategic Command on 1 October 2002. This combined Command assumed current space and strategic nuclear deterrence missions of its two component commands and also global strike and information operations missions. The latter two missions almost certainly would involve space capabilities (5).

It is unclear how these future force application missions might be controlled. However, several aspects of this control are worth noting. First, they would probably be coupled with real-time targeting systems, also based in space, to track their missile or other moving air, maritime, or ground targets. The U.S. Space-Based Infrared System (SBIRS) now under construction is one such system. The proposed Discover II system, which was under development before it was cancelled in 2000, would have used several dozen space-based radar satellites to track moving targets on the ground. In addition to these real-time surveillance and tracking systems, a command and control system that had very high fidelity and rapid response would also be needed. Engagements in and through space occur on timescales of seconds, so it is unlikely that a future command and control system would use human operators for conducting detailed engagements—there is simply insufficient time for human reactions to meet the requirements. Conversely, future force application systems would undoubtedly require “man-in-the-loop”—not as noted before to conduct the engagement, but to enable the system and set operational parameters (limitations) at the beginning of a crisis or engagement. Moreover, for reasons of survivability—particularly for critical strategic systems such as missile defenses, the command and control system would undoubtedly be “distributed,” rather than confined to a single or small number of potentially vulnerable nodes.

Other future missions may also emerge. In recent years, there have been growing calls and concerns about the threat of natural objects striking the earth—comets and asteroids. A large asteroid strike is generally deemed responsible for wiping out the dinosaurs—along with most other large animals—65 million years ago. And small asteroids have apparently struck the earth with explosive power comparable to a nuclear weapon several times a century. It is uncertain what might be done about these threats, but most agree that a comprehensive all-space surveillance system is needed to catalog potential threats and track objects as they move close to Earth. This capability is similar to that needed for controlling and monitoring the growing constellations of man-made satellites and almost certainly will be an adjunct mission of any military space-based space surveillance system. Should a threatening asteroid or comet be detected, there have been various proposals to divert the threat. Many of these proposals include the use of a nuclear weapon exploded in the vicinity of the threat to deflect or destroy it. This would invariably require international cooperation but could also include various military capabilities and systems such as launch and space intercept systems.

The possibility of growing national security reliance on space has raised, at least in the United States, the potential need for a military service, separate from today's army, navy, and air force. Russia has recently established such a separate

space service adding credibility to these proposals. The U.S. Congress sponsored a high-level commission in 1999 to study this possibility. Donald Rumsfeld, who was subsequently appointed U.S. Secretary of Defense, chaired it. The commission recommended that a separate space service was not appropriate at the current time but was likely to emerge in the decades ahead.

“Space Commission” Changes. The charge of the aforementioned commission was to “Assess the organization and management of space activities that support U.S. national security interests” (26). After nearly a year of deliberation and interviews with nearly every senior member of every organization within the U.S. government responsible for some aspect of space development and operations, the Commission’s findings resulted in some significant changes in the DOD’s structure for spacecraft and space launch life-cycle system oversight and management. Their overarching findings were that it was “in the U.S. national interest to

- promote the peaceful use of space;
- use the nation’s potential in space to support its domestic, economic, diplomatic and national security objectives;
- develop and deploy the means to deter and defend against hostile acts directed at U.S. space assets and against the uses of space hostile to U.S. interests” (26).

The Commission further recommended a full review and revision of the nation’s space policy. Ultimately the commission recommended that the policy provide unambiguous direction to all branches of government to ensure that space systems developed address the needs of the military to deter and defend against evolving threats to its forces, allies, and other international interests. The commission further identifies the changes that will have to be made to make this possible. They include revolutionary use of space for collecting and disseminating intelligence to facilitate effective planning and resolution of national crises; development of an international legal and regulatory environment for space issues that ensures U.S. national security interests and enhances commercial competitiveness and effective exploitation of space for civil purposes; and finally, the commission recommended renewed investment by the government and commercial sector in leading edge, truly revolutionary, technologies to ensure that the United States can master operations in space and continue to compete on the open market. Of particular interest to the DOD were the recommendations the commission made with regard to the structure of the military’s organizations involved with acquisition and operation for and in space.

From the commission’s research and through the many interviews they held, the commission quickly realized that the Intelligence Community and the Department of Defense “are not yet arranged or focused to meet the national security space needs of the 21st century” (26). In response to this shortcoming, the commission concluded that the numerous space activities throughout the Defense and Intelligence Communities should be merged and have chains of command, lines of communications, and policies adjusted to ensure improved

accountability and responsibility. To this end, the DOD has initiated several notable changes in its management structure for space.

In April 2002, the U.S. Air Force Space Command (AFSPC) was assigned its own four-star commander. This change will ensure that AFSPC has the necessary autonomy to assert authority appropriately over its new organization. As the newly appointed executive agent for space, AFSPC is now the leader for cradle to grave development and deployment of military space systems. In support of this designation as executive agent for space, responsibility for SMC has been transferred from the Air Force Material Command to AFSPC. The net result of the changes made is yet to be realized; however, this marks the first time that one organization has been responsible for military space systems. This, along with their recommendation to consolidate budgeting for space programs, will ensure that the DOD is able to address the requirements of its forces and ensure that U.S. national security interests are appropriately addressed.

Summary

Most space programs throughout the world emerged from national security programs. Generally, national security use of space for communications, surveillance, and other information is the first practicable use for space that a nation sees. This was true, particularly for the United States. National security use of space continues to be the lion's share of U.S. government space investment. Although there is a growing trend to rely on commercial and dual use civil and military space infrastructure—particularly for space launch support, unique military space control and operations facilities are here to stay. As various nations, probably led by the United States, make even more use of space for national security, these unique military operations facilities will evolve considerably. They are likely to become much more “distributed.” They will increasingly rely on space-based elements such as space-based space surveillance systems and space-relayed command and control systems. It is even possible, if not likely, that we will see national security space forces emerge separate from terrestrial armies, navies, and air forces.

BIBLIOGRAPHY

1. Lasker, L., and W.F. Parkes. *War Games*. MGM, 1983.
2. Chapman, R.G. Jr. (for U.S. Air Force Space Command). *Legacy of Peace: Mountain with a Mission, NORAD's Cheyenne Mountain Combat Operations Center*. The New Mexico Engineering Research Institute University of New Mexico, Albuquerque, NM, 1986, pp. 1–3.
3. National Reconnaissance Office. Who We Are, Available WWW: <http://www.nro.gov/index1.html>.
4. Defense Technical Information Center. Goldwater–Nichols Act, Available WWW: http://www.dtic.mil/jcs/core/title_10.html.
5. U.S. Department of Defense. DOD Announces Merger of U.S. Space and Strategic Commands. News Release, 26 June 2002, Available WWW: http://www.defenselink.mil/news/Jun2002/b06262002_bt331-02.html.

6. Report of the Commission to Assess U.S. National Security Space Management and Organization. Washington, DC, 2001. Also available on WWW: <http://www.defenselink.mil/pubs/space20010111.html>.
7. Garamone, J. *U.S. Northern Command to Debut in October*. American Forces Press Service, U.S. Dept of Defense, 17 Apr 2002. Available WWW: http://www.defenselink.mil/news/Apr2002/n04172002_200204175.html.
8. Launius, R.D. *Organizing for the Use of Space: Historical Perspectives on a Persistent Issue*. Univelt, San Diego, CA, 1995.
9. HQ AFSPACECOM/HO. Aerospace Defense: A Chronology of Key Events, 1945–1990. 1991.
10. Valentine, L.Y., and M.A. Chevalier. Air Force Space Command: The Best Space and Missile Team in the World. AFSPC Affairs, Available WWW: <http://www.spacecom.af.mil/hqafspc/library/almanac/cover/cover.htm>.
11. Army Space and Missile Development Center History. Available WWW: <http://www.smdc.army.mil/CmdHistory.html>.
12. Naval Space Command: Established to Consolidate Sea Services' Space Operations. Available WWW: <http://www.navspace.navy.mil/history/history.htm>.
13. Darobeck, C. Navy Creates Network Command. Federal Computer Week, 27 Mar 2002, Available WWW: <http://www.fcw.com/fcw/articles/2002/0325/web-navy-03-27-02.asp>.
14. USAF Space and Missile Systems Center. Historical Overview 1954–1995 Space & Missile Systems Center. Available WWW: <http://www.losangeles.af.mil/SMC/HO/Smchov6.htm>.
15. HQ NORAD/HO. Cheyenne Mountain Operations Center: History, Available WWW: <http://www.cheyennemountain.af.mil/today.htm>.
16. Hanley, T.C., and H.N. Waldron. Historical Overview: Space & Missile Systems Center 1954–1995—Satellite Systems. SMC, Los Angeles AFB, CA. Available WWW: <http://www.losangeles.af.mil/SMC/HO/Smchov.htm>.
17. National Oceanic and Atmospheric Administration. National Polar-Orbiting Operational Environmental Satellite System. Available WWW: <http://www.ipo.noaa.gov/gallery4.html>.
18. HQ/AFSPC. Air Force Space Command Fact Sheets: Navstar Global Positioning System. Available WWW: <http://www.spacecom.af.mil/hqafspc/Library/FactSheets/FactsSheets.asp>.
19. Defense Information Systems Agency. Defense Information Systems Network (DISN) Architecture. Available WWW: <http://www.disa.mil/DISN/disnarch.html>.
20. HQ/AFSPC. Air Force Space Command Fact Sheets: Defense Satellite Communications System. Available WWW: <http://www.spacecom.af.mil/hqafspc/Library/FactSheets/FactsSheets.asp>.
21. Naval Space Command. Satellite Communications Enable Information Technology for the Twenty-First Century (IT-21). Available WWW: <http://www.navspace.navy.mil/NEWS/IT21.HTM>, 8 July 1999.
22. Naval Space Command. Navy Takes Control of Communications Satellites. Available WWW: <http://www.navspace.navy.mil/NEWS/UFO.HTM>, 21 Oct 1999.
23. Ray, J. *MILSTAR Satellite Launched in First Space Shot of 2002, Spaceflight Now*. Polestar 2002. Available WWW: <http://spaceflighnow.com/titan/b38>.
24. USAF Air University. Space Surveillance. Available WWW: <http://www.maxwell.af.mil/au/awc/awcgate/usspc-fs/space.htm>.
25. Aldinger, C. Laser Hits Its Mark in Tests by the Pentagon. Reuters News Service, 21 October 1997, Available WWW: <http://www.freep.com/news/nw/qlaser21.htm>.
26. Surrey's Nanosatellite Technology. Surrey Technology Limited. Available WWW: http://www.sstl.co.uk/services/subpage_services.html.

READING LIST

- Canadian Army Headquarters. A Brief History of the Canada-United States Permanent Joint Board on Defence 1940–1960. The Queen's Printer, Ottawa, Canada, 1960.
- Cheyenne Mountain Study*. ANSER Corporation, 1992.
- Headquarters, U.S. Air Force. Briefing on the United States/Canada Joint Surveillance System. Washington D.C., HQ USAF/CVAI I Internal Affairs Division, Office of the Vice Chief of Staff, 29 April 1983.
- North American Aerospace Defense Command, NORAD's Underground COC—Initial Requirement to Initial Operation 1956–1996, NORAD, Colorado Springs, CO, 30 December 1985.
- Lederman, G.N. *Reorganizing the Joint Chiefs of Staff: The Goldwater-Nichols Act of 1986*. Greenwood Press, Westport, CT, 1999.
- Osgood, J. The Goldwater Nichols Act - Managing the Defense Department. Available WWW: <http://pw1.netcom.com/~jrosgood/w16.htm>.

ADAM L. MORTENSEN
SIMON P. WORDEN
USAF, USSPACE/SIOE-r
Colorado Springs, Colorado

MILITARY USE OF SPACE

Starting at the very beginning and continuing until now, one of a nation's objectives for the conquest of space is to obtain military strategic advantages by using such unique features of space as extraterritoriality, global access, unique physical conditions, and a number of others. Thus, in the late twentieth century, the revolutionary development of humanity's new technological capacities inexorably moved the world into a new phase of geopolitical confrontation — the race to attain strategic advantages in space, which, in turn, made military operations in space one of the most important ways to ensure national security. This implies that the significance of various arenas of confrontation will change in the twenty-first century; space will grow considerably in importance as the arena for attaining the objectives of armed conflict (Fig. 1). The world's leading nations thus accord space a special position in contemporary military doctrine and approaches to national security.

The "The United States' National Security Strategy for a New Century" (1998) expresses Washington's principal approaches to the use of space in the interests of national security. According to this document, the United States intends to maintain its leadership in space, attain unimpeded access to space and to its use in the interests of protecting national security and ensuring the well being and prosperity of the nation. The fact that space has acquired importance from the standpoint of access to global information is stressed, because this has serious political, diplomatic, military, and economic consequences for the United States. Space is seen here as one of the major arenas in which current and future technologies can play a key role in the use of military power for generating a

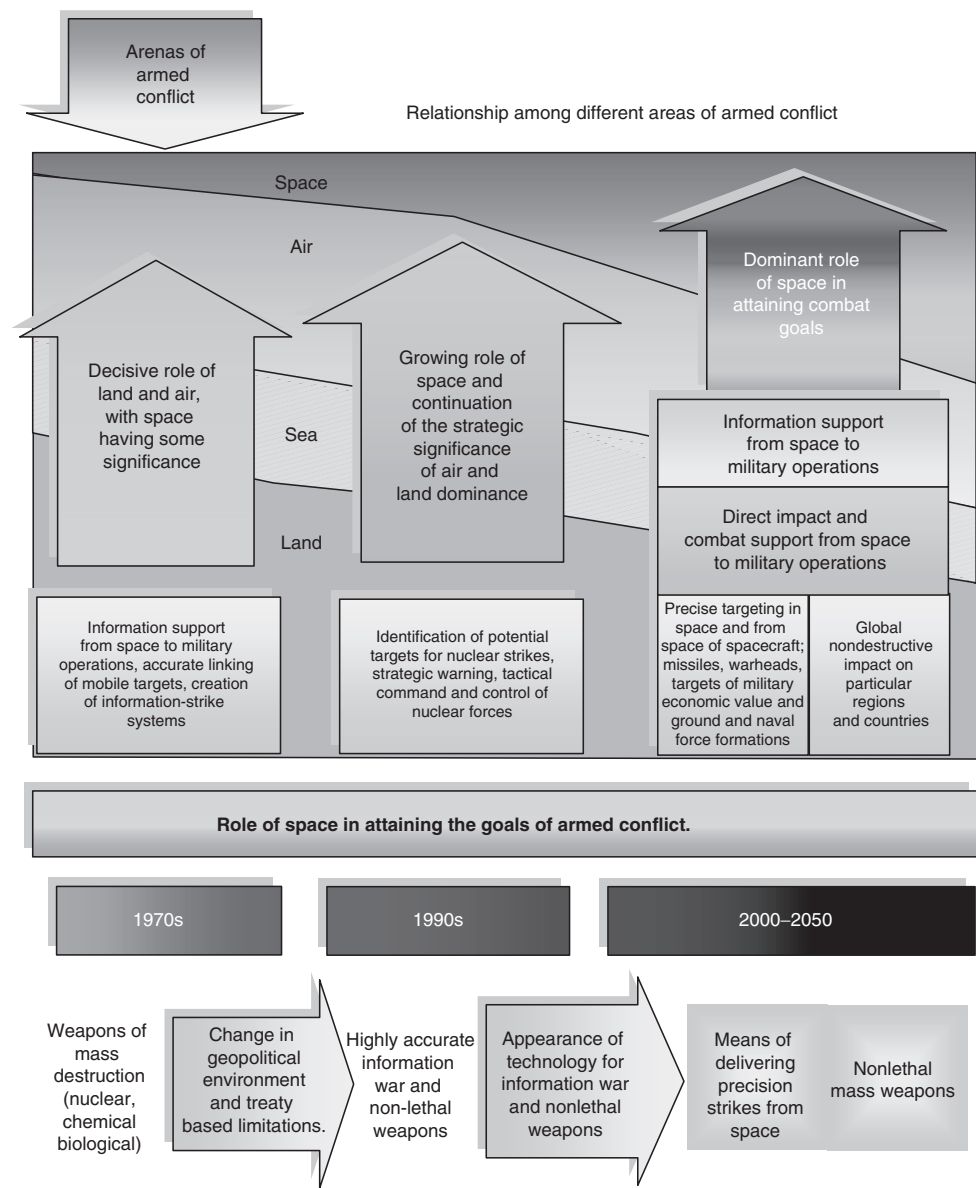


Figure 1. Change in the significance of arenas of armed conflict. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

favorable international climate and in the capacity to react appropriately to the entire range of potential threats and crisis situations.

Russia’s national security is considered comprehensively and in depth in “Concepts of National Security” adopted by the Russian Federation President’s Edict No. 24 on 10 January 2000. The major thrust of this policy is directed at creating a multipolar world by strengthening the political positions of a

significant number of states and international associations, improving the mechanisms of multilateral control of international processes, and expanding economic, scientific, and technical collaboration, including collaboration in space.

Note that all phases of the conquest of space were, to some degree, associated with using space technology to solve practical military problems. Such major problems have traditionally included

- prevention of missile and space attacks;
- reconnaissance and informational support;
- navigational and geodesic support;
- meteorological support;
- communications among military users.

As a result of the arms race, starting in the mid-1970s, this list of tasks was augmented by a completely new thrust associated with the development of space-borne weaponry, and space began to be considered a potential arena of armed combat.

Retrospective analysis shows that the scale of military use of space has had a tendency to expand and has gone through a number of phases (Table 1). The foundation for the military use of space was laid virtually immediately after it became possible to insert spacecraft into near-Earth orbit (1957–1959). During this period, experimental spacecraft were launched into space, and the key issues of deploying satellite constellations, controlling them, and obtaining information from space were formulated.

In 1961, a new phase in the military use of space began—the first special purpose military photoreconnaissance satellite, Zenith, was launched. For two years more than 10 similar satellites were launched, after which the first space electronic intelligence complex entered service.

During the period between 1966 and 1976, more advanced Zenith photo-reconnaissance satellites, as well as electronic intelligence systems (Tselina, US-P), radar reconnaissance systems (US-A), positioning systems (DC type), geodesic satellites (Sfera), space communication systems (Molniya, Strela), weather satellites (Meteor), and navigational satellites (Tsikada, Parus, and others) were flights-tested and adopted. Also during this period, development of space-borne missile defense systems was begun and, between 1972 and 1976, four experimental satellites of this type were launched (type US-K).

The development and use of space technology abroad (primarily in the United States) began during the same period and followed a similar course. Thus, the first experimental reconnaissance satellite, Discoverer-1, was launched on 28 February 1959. Spacecraft in this series were used to develop means and methods of space reconnaissance. In 1960, the United States began to use the Samos series spacecraft for imaging reconnaissance. There were three generations of Samos vehicles (Samos-2, D1, Samos-P and Samos-M). Other satellites launched during this period included the Ferret electronic intelligence spacecraft, the Score and Syncom communications satellites, and the Tiros military weather satellite. Space-based missile attack warning systems (originally Midas, then IMEWS) and systems for detecting ground-based nuclear explosions flown

Table 1. Phases in the Use of Space for Military Purposes

1959	1961	1976	1991–2006
Experimental research	Solution of individual military problems	Full-scale informational support of military forces from space	Operational equipping of space as a new arena of armed combat
Experimental launches of developmental spacecraft	Launches of individual spacecraft	Permanent space systems	New generation space systems integrated with armed services weapons systems
Formations of 1–2 spacecraft	Formations of 5–10 spacecraft	Formations of 100–120 spacecraft	Formations of 150–200 spacecraft
Problems solved: *development of space-missile complexes and onboard spacecraft systems	Problems solved: *individual problems in photo and electronic reconnaissance; *individual problems related to navigation, communications, geodesics, and positioning	Problems solved: *missile attack warning; *photo-, electro-optical, electronic, and radar reconnaissance *communications and tactical command and control, geodesic, weather, and cartographic support; *positioning; *provision of remote sensing information for strategic and operational-tactical levels of command and control	Problems solved: *missile attack warning; *real-time, global, all-weather, and target designation *support of communications and command and control *navigational, topographic, meteorological, and cartographic support *positioning *monitoring of arms reduction treaty compliance *use of space technology for controlling weapons systems of different services *provision of information from space for strategic, operational-tactical and tactical (on the battle field) levels of command and control in space and from space *development of networks of small spacecraft *development of technology for combat in space and from space

on the Vela spacecraft at high (110,000 km) circular orbits were considered particularly important. During this period, a communications system was deployed in geostationary orbit, and the United Kingdom (Scynet-1A) and Canada (ISIS-1) developed their own spacecraft.

During the 1970s, the United States developed and deployed the more advanced LASP and then KH series reconnaissance spacecraft, making it possible to conduct wide area and detailed surveillance. The Riolite satellite, which had a large antenna for radiointerception over Europe, was inserted into geostationary orbit and the space-based missile warning system and space based communication, navigational, and weather systems were improved.

As a result of the development of new generation space systems and technologies with considerably increased active life spans and better onboard equipment and data transmission systems, there was a qualitative improvement in the use of space technology to solve military problems. Permanent orbital formations of space-borne systems for various purposes were deployed to provide informational support for the operations of the armed services. The number of problems that were being solved by using space technology (Table 2) and the contribution to support of military operations (Table 3) increased substantially. As a result, extensive use of space technology became generally accepted and expected when planning the strategic operations of the armed forces and in planning operations of ground and naval forces at the operational level, as well as of tactical formations. The technological capabilities to build means of space combat and implementation of work to do so laid the foundation for identification of space as an independent arena for military operations.

A new round of space militarization started in the 1980s after President Ronald Reagan's famous "Star Wars" speech, after which the United States developed the Strategic Defense Initiative (SDI) program, which for nearly 10 years, had a decisive influence on the formation of the U.S. administration's national space policy. The proclaimed objective of this program was defense, primarily of U.S. territory, but also that of their allies, against nuclear strikes.

In the course of work on the SDI program, many projects were proposed for space technology equipped with various types of weapons (Fig. 2). One of the projects of this type was a system based on several thousand small interceptor spacecraft (Fig. 3). All of these projects dealt with concepts of weapons for the remote future and were not embodied in specific models or plans to deploy new systems of space-based weapons. Aside from such projects, specific steps were taken to create space weapons, for example, flight tests of the ASAT antisatellite system, by which a missile with an infrared homing device, launched from an F-15 aircraft, was designed to destroy a spacecraft.

Under these conditions, the Soviet Union too began work to counteract this space militarization program, in the event it was implemented. Thus, anti-antimissile defense projects began to be developed. These involved the joint use of passive and active means of defense, mounted directly on an ICBM to create a "launch window" during a nuclear rocket strike by disrupting the space grouping of antimissile defenses, creating information gathering devices for real-time reconnaissance of the space environment and providing target designation data to weapons to combat the space interceptors, and developing space-borne weapons capable of striking targets on U.S. territory and at sea. The implementation of

Table 2. Problems Solved Using Space Technology

Purpose of space technology	Principal military/political phases		
	Peacetime	Period of threat	Initial and subsequent periods of war
	Objectives of using space technology		
Missile threat warning	Supporting the day-to-day operations of ground and naval forces	Detecting signs that the enemy is preparing to attack and providing information for planning combat operations	Supporting ground and naval forces for planning combat operations and the use of weapons
	Detecting of the launch of ballistic missiles and warning of missile attacks		
	Operational reconnaissance of space environment and generation of data for target designation Identification (refinement) of deployment of military forces, determination of their characteristics and coordinates Disclosure of signs of change in the makeup, location, and level of combat readiness, observation of regions of local wars and large training exercises		
Reconnaissance	Monitoring compliance with arms reduction agreements and treaties		
		Observation of regions with electronic facilities and launch sites, exact determination of their coordinates, provisions of data to agencies in charge of reconnaissance, weapons, and electronic warfare	
		Surveillance of strike, defensive, and supporting formations of naval forces and provision of data for target designation to naval weapon systems Discovery of measures relating to operational outfitting of combat regions, regrouping of strike forces and reserves	
Communications, tactical command, and control and relay	Support of command and control of strategic nuclear forces Support of communications and data transmission in command and control of armed forces Relay of reconnaissance data from space		Monitoring of the results of weapon employment
	Provision of data for navigational location of moving objects of ground and naval forces		
	Collection of weather data for agencies providing command and control of forces and weapons; issuing of weather forecasts and climate information		
Navigation	Provision of data for making and updating topographic and digital maps, city plans, and photographic documents	Refinement of topogeodesic descriptions of combat regions	
Meteorology			
Cartography			
Geodesy	Provision of data for refining geodesic constant parameters of the terrestrial ellipsoid and navigational field of Earth		
Special support	Provision of alignment and calibration of air and space defense systems		

Table 3. Contribution of Space Systems and Technologies to the Solution of Problems of the Armed Forces During Different Military and Political Periods

Problems solved by space systems	Results of use of space forces and technology (qualitative indicators)			
	Peacetime	Local conflicts	Conventional large scale war	Nuclear war
Missile attack warning	More accurate monitoring of regions with dangerous missile launch capability in nuclear nations (in the future all countries that have missile potential) and of the ocean			More accurate detection of BM launches, increased accuracy of flight path prediction
Strategic reconnaissance	More accurate monitoring of compliance with treaties, detection of signs of preparation for war, measures to outfit theaters of war operationally and identification of the makeup, deployment, and combat readiness of armed forces			
		More accurate monitoring of enemy attack forces, evaluation of intentions and direction of the main strike		
Operational and tactical reconnaissance	More accurate monitoring of state of enemy facilities	More accurate estimation of results of weapon use		
Communications and command and control	More reliability, interference resistance, and throughput for command and control systems; increased efficiency in getting reconnaissance information to the troops			
Navigation	Increased safety of aircraft and ships	Increased efficacy in destruction of enemy facilities, formations, and weapons Increased capabilities for use of aircraft in flight, ships and submarines in the ocean, formations of troops and mobile forces in any region, even when they are not armed		
Geodesic support	Increased support of forces with geodesic information on enemy territory for the use of the SNF	Increased accuracy of topogeodesic descriptions of combat regions		
Meteorological support	Increased provision of weather data to forces and strategic weapons			

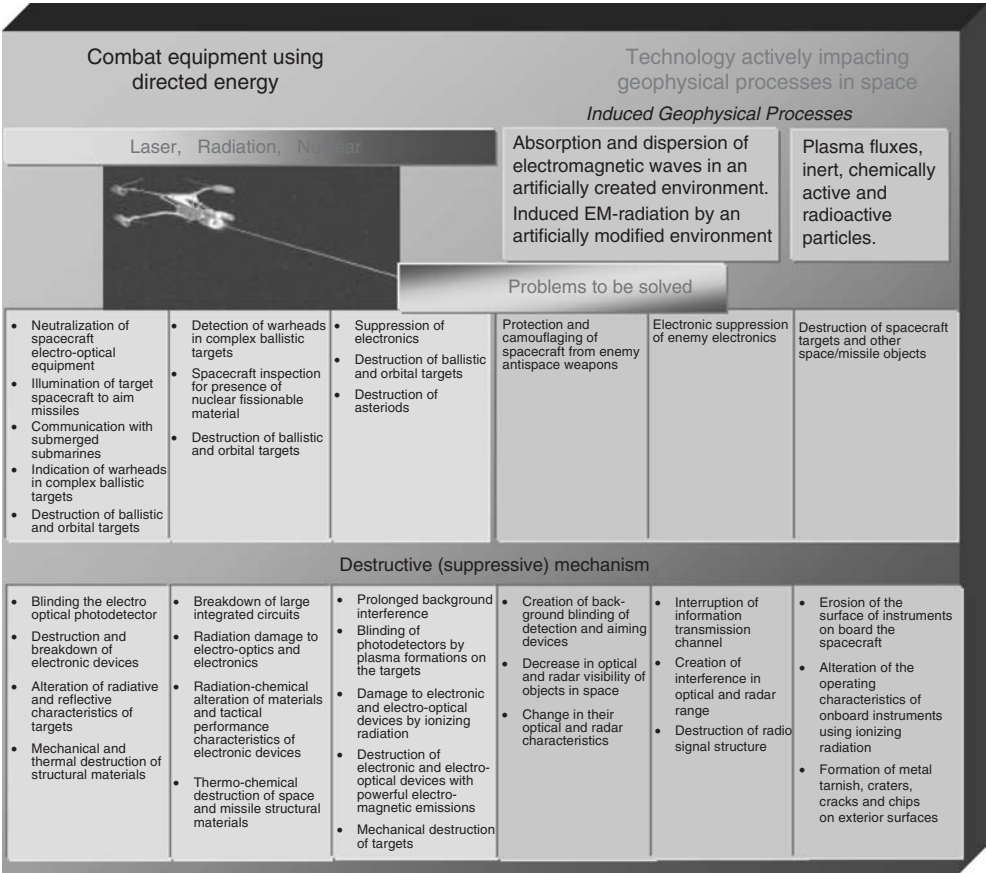


Figure 2. Projected advanced combat equipment. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

these anti-antimissile defense measures would have made it possible to penetrate the “space umbrella” over U.S. territory sufficiently to deliver unacceptable losses. Furthermore, in response to tests of the ASAT system, launches of the “IS” system in the Soviet Union were extended (Fig. 4) as it was in constant use, and various projects were initiated to develop space warfare technology.

All of this led, at the start of the 1990s, to a shift from the theoretical aspects of investigating space as a new arena for armed combat to solving operational problems of equipping near-Earth space as a possible theater of military operations (by analogy to the continental and naval theaters), in combination with the deployment of components of space infrastructure on the ground for launching spacecraft (combat and remote sensing), ground-based means of controlling them and receiving information to support a full spectrum of warfighting use of space systems and technology. One possible schematic depiction of space as a strategic space zone is presented in Fig. 5.

Whether space plays a dominant role in attaining warfighting objectives at this time will be determined by the extent to which it is possible to develop and deploy space-based weapons technology for conducting military operations in and

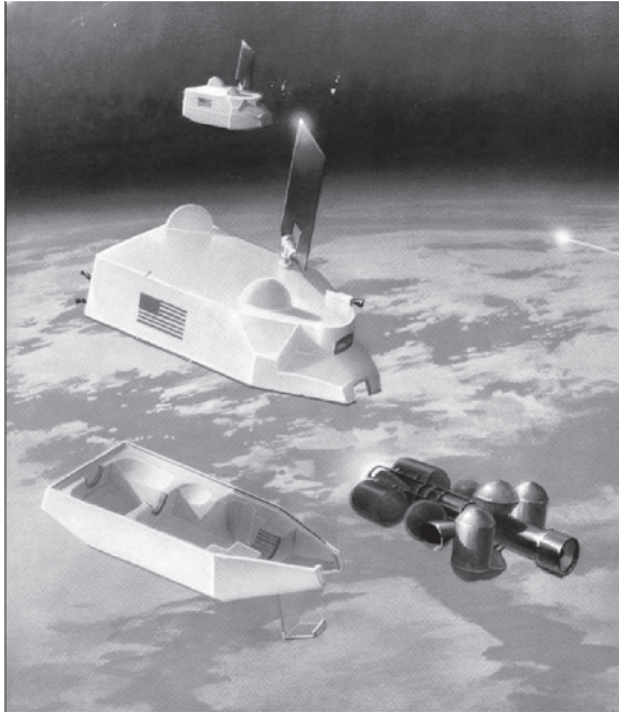


Figure 3. Orbital formations of U.S. antimissile Brilliant Pebbles (BP) spacecraft. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

from space that provide both strike capacity and combat support. This is the reason that the leading space powers are developing directed energy and kinetic weapon systems for destruction of targets. Plans call for using these weapons against ground-based facilities and aircraft in combat. An indicator of the growing significance of space combat forces is their inclusion, along with nuclear missiles, in the Strategic Command.

The growing role of these systems is underscored by the fact that the United States, having announced its withdrawal from the 1972 Anti-Ballistic Missile Treaty, is actively implementing a program to develop missile defense systems, whose weapon and information components will have a significant space component.

On the whole, the early 1990s were marked by headlong growth in the space potential of the world's leading powers (primarily U.S. and Russia) and qualitative changes in the use of space technology to resolve military problems and issues of national security. Although previously space technology was generally used only episodically in local wars and armed conflicts (Vietnam, the Near East, Afghanistan, Falkland Islands, etc.) and was a function of the presence of a certain reconnaissance satellite in orbit and whether or not it would pass over a given surveillance region at a particular time, by 1991, one can speak of wide-scale practical use of space systems in the course of military operations. This was demonstrated in the Gulf War, when the multinational forces used space technology in all phases of their operations against Iraq. The main problems facing those

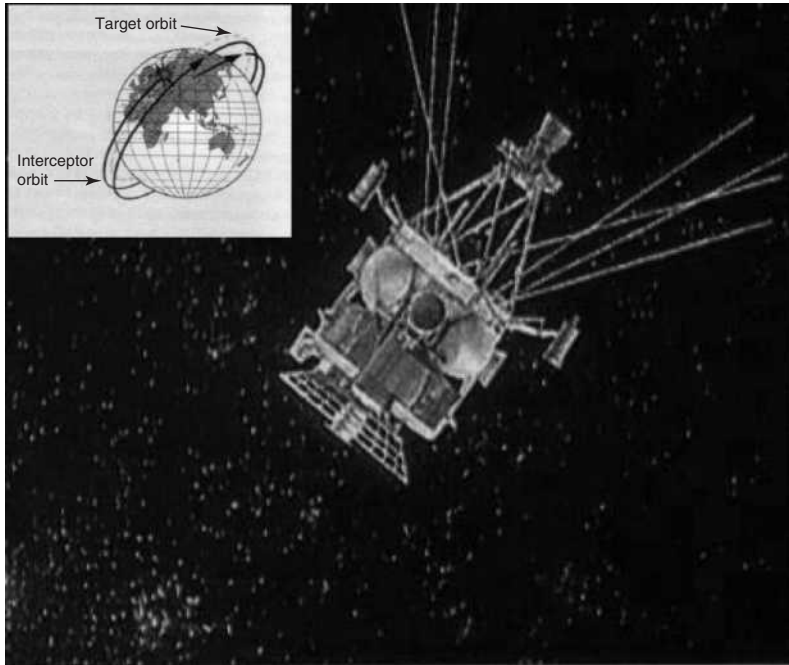


Figure 4. Exterior view of the IS spacecraft and a scheme of target interception in space. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

charged with the control of the space systems in the conflict region involved supporting reconnaissance and communications, damage assessment of enemy targets; and navigational, topogeodesic, and meteorological support of the forces.

On the whole, space technologies had such a strong effect on the operations of the multinational forces in the Persian Gulf conflict that new tactical methods for using them in warfare were developed. According to the experts, the Gulf War was the “first war of the space era” or the “first space war of our era.” The application of space technology, specifically the use of remote sensing information was even more impressive in scope in the military operations of NATO forces in Yugoslavia. During this conflict, the creation of a comprehensive reconnaissance system using aircraft and space was virtually completed. This system enabled real-time recovery of both strategic and tactical information. In the United States the space-based portion of this system included the KH-11, 12 electro-optic reconnaissance satellites, the Lacrosse radar reconnaissance satellite, the Magnum and Vertex DMSP weather satellites, and also the French SPOT satellites. The aircraft component consisted of manned and unmanned reconnaissance planes of the Hunter, SD-289, and Predator types. The data obtained were transmitted to a national processing center in Washington, D.C., or to information reception stations in the conflict area (ground stations of the Constant Source types and shipboard terminals of the FIST type)—Fig. 6.

Installation of receivers of Navstar navigational signals on moving objects (including integrating them into onboard control systems of high-accuracy weapons) made it possible to link objects very precisely to a supporting information

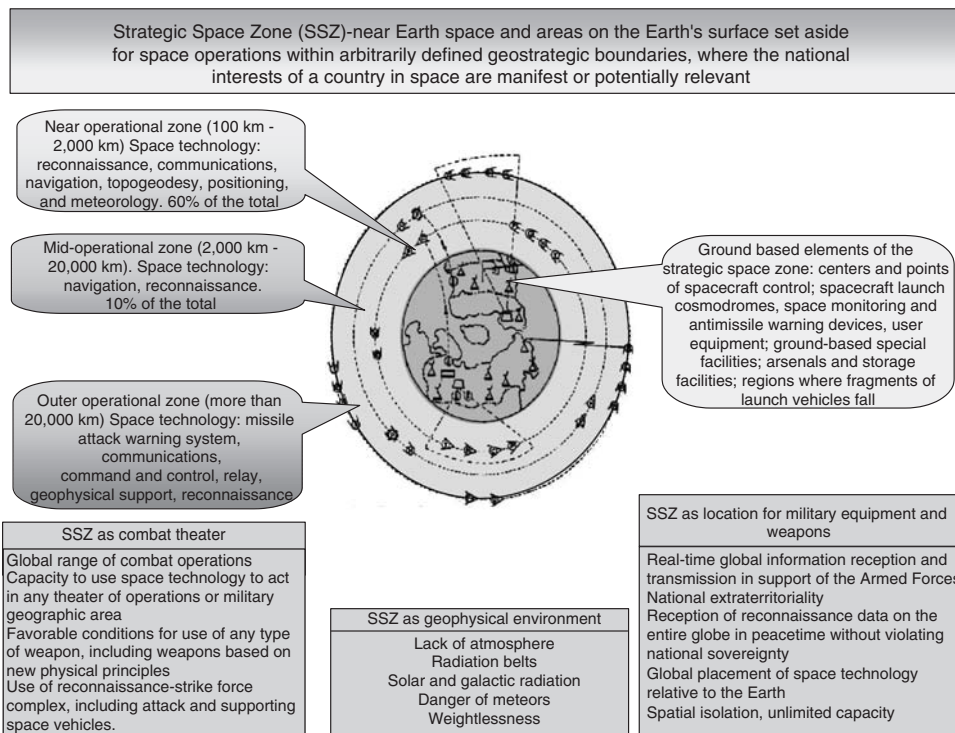


Figure 5. Space—the Strategic Space Zone. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

field, and thus, to hit targets in Yugoslavia, in a number of instances without NATO aircraft having to enter the air defense zone.

The significant development of the mechanism for organizing the use of space communications and data transmission systems enabled real-time creation of a high-bandwidth adaptive communications and data transmission network in the conflict region. This, in turn, led to increased “depth” of use of space-based communications, making it possible to tailor the flow of data about combat conditions and to organize joint use of space systems belonging to different nations efficiently.

Note that the use of civilian-developed and commercially available systems also played an important part in solving a wide range of applied military problems through

- development of mechanisms for obtaining civilian and commercial space systems for solving of military problems (thus, civilian systems are widely used by military agencies by renting channels on commercial communications satellites; the U.S. Department of Defense also gets a large amount of information from civilian satellites that surveying Earth’s natural resources, geodesy, and weather and uses more than 20% of the information obtained from the U.S. Landsat system, supplementing it with information from the SPOT (France) and MOS (Japan) remote sensing satellites; the cartographic

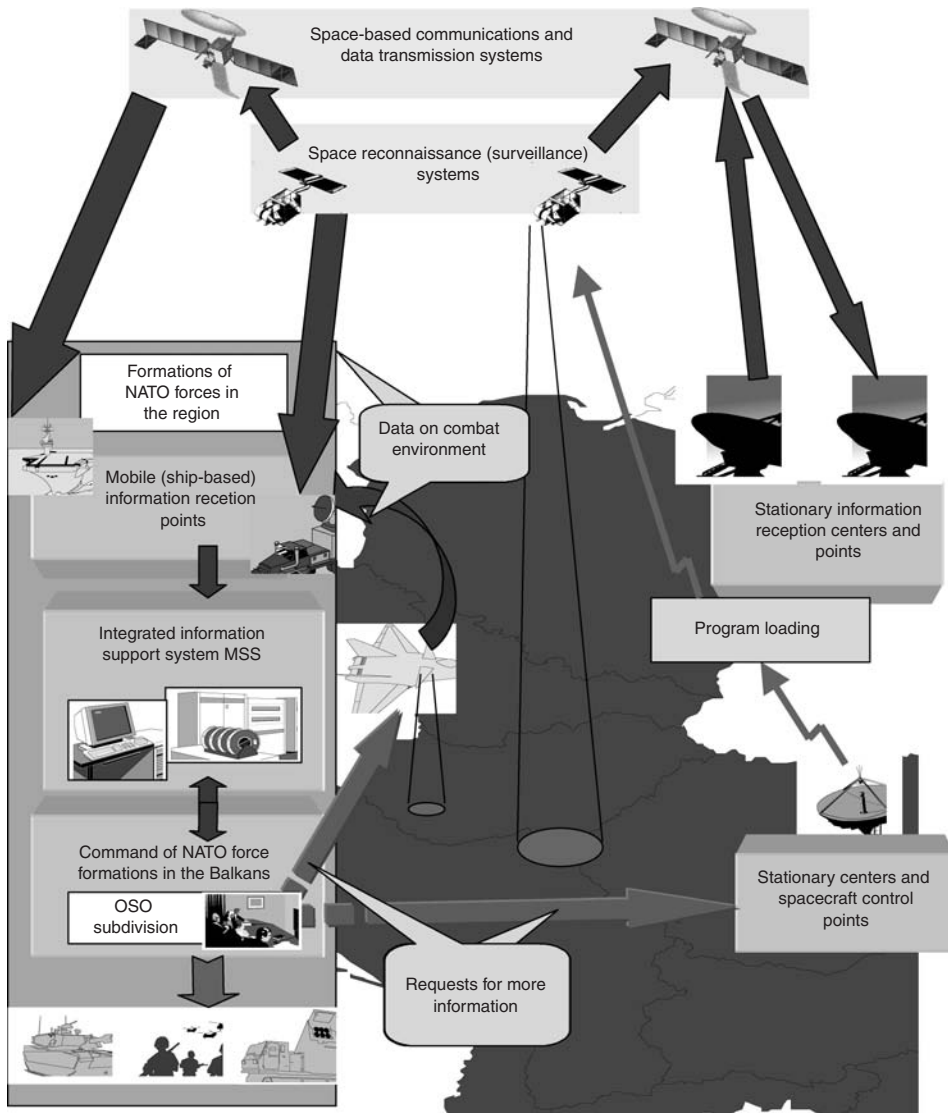


Figure 6. Functional scheme of the complex aerospace reconnaissance system. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

administration of the U.S. Department of Defense is second only to the Department of Agriculture with respect to the number of photographs purchased from natural resource surveillance satellites);

- organization of interactions among the leading military and civilian agencies coordinating new technology development (DARPA, NASA, and others) in the form of joint projects (TRP project) and bilateral agreements for coordination of work in the area of new technologies (agreement between NASA and the Air Force Space Command, signed in February, 1997).

The development of space technology for solving analogous problems in Russia has proceeded in three interrelated directions. The first direction involves the development of space technology to serve the needs of wartime. The technical basis of this developmental trend is work to develop small spacecraft and launch devices. This direction is based on the transition to a new scientific and technical stage distinguished by significant miniaturization and increased reliability of electronic technology. The development and deployment of systems based on small spacecraft entails their use to solve what are mainly tactical problems. Plans call for using these systems to support communications within the limits of a theater of operations, for observing enemy troop movements, for obtaining data for damage assessment, and for performing certain experiments.

The second direction involves development of systems for bringing remote sensing information to the lowest level of troop command. This direction only began to be developed toward the end of the twentieth century, when models of powerful, small sized artificial intelligence technology began to appear and changed the very understanding of the nature of contemporary warfare. During the implementation of this direction, positive experience was gained using the specialized Space Support Group at operational and tactical levels of command. The essential tasks of these groups are evaluating the status and capacity of the spacecraft and preparing commands to obtain data and relaying the information received (reconnaissance, weather, navigation, and communications) to the commanders at various levels of command with recommendations for their use. The tasks that the Space Support Groups must perform require that they include (Fig. 7) coordination groups for using space capabilities and reconnaissance technology, communications, navigation, and meteorological support, and also groups for processing and integrating the remote sensing information obtained. To operate efficiently, the Space Support Groups require specialized mobile stations for receiving and processing information from satellites and then issuing it in a convenient form for the user. One way such a station can be structured is presented in Fig. 8.

The third direction involves development of dual-purpose systems that solve the problems of both military and civilian users based on

- radical improvement of the ground-based space infrastructure (adaptation of the spacecraft control systems; creation of highly efficient launch facilities and mobile control systems);
- adoption of advanced spacecraft control technologies based on space-based relay and navigation systems;
- development of highly specialized dual-purpose spacecraft systems using onboard information processing based on state-of-the-art information technologies;
- optimization of configurations of dual-purpose space systems.

Implementation of these developmental directions has raised the process of force command and control to a qualitatively new level and has multiplied their combat potential severalfold. This will require corresponding qualitative changes in

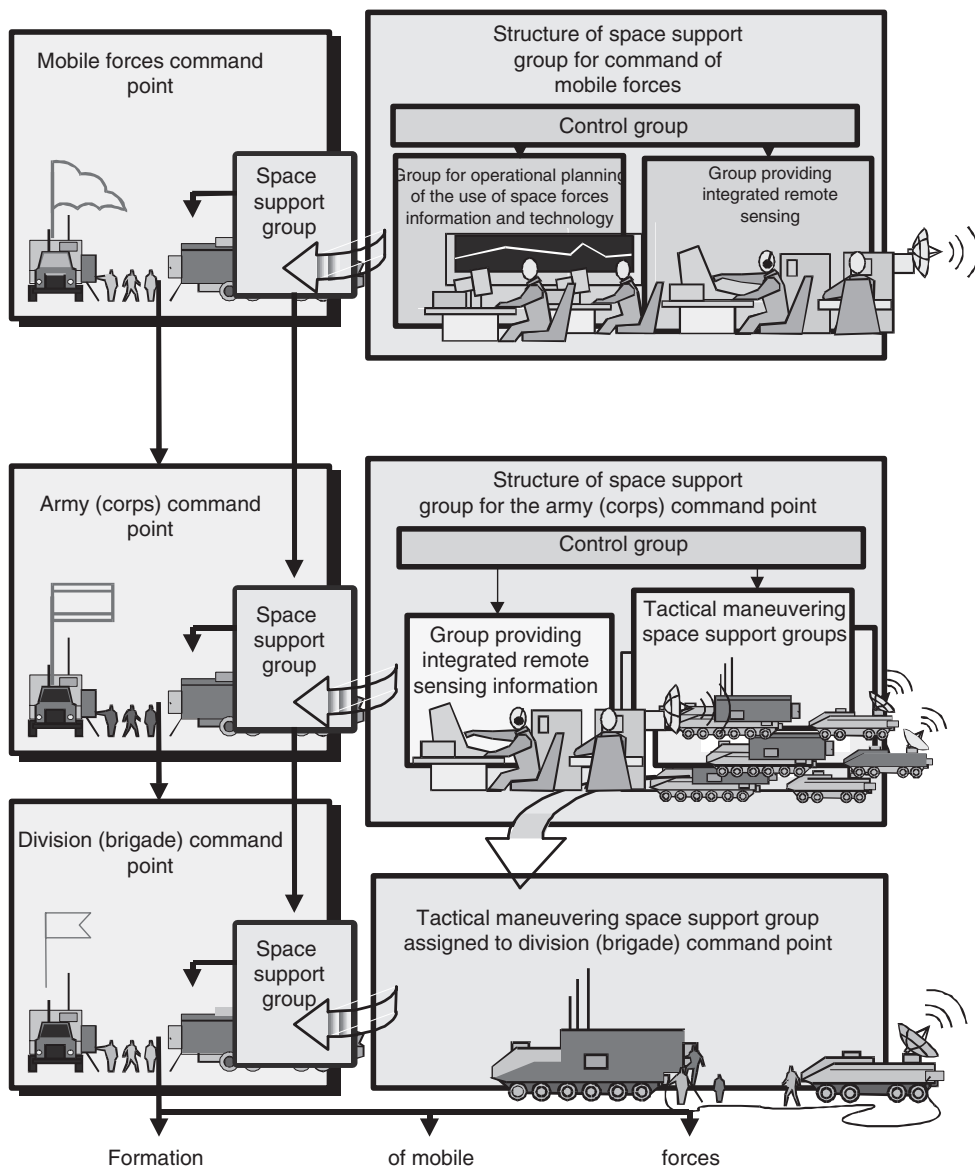


Figure 7. One version of the structure of space support groups within mobile force command points. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the design of spacecraft and command systems, the technology for evaluating and depicting remote sensing information, and support of decisions based on such information.

In addition, the development of information technologies and the start of the era of information warfare (information competition) has created the need for the next step in the applied military use of space, which will involve

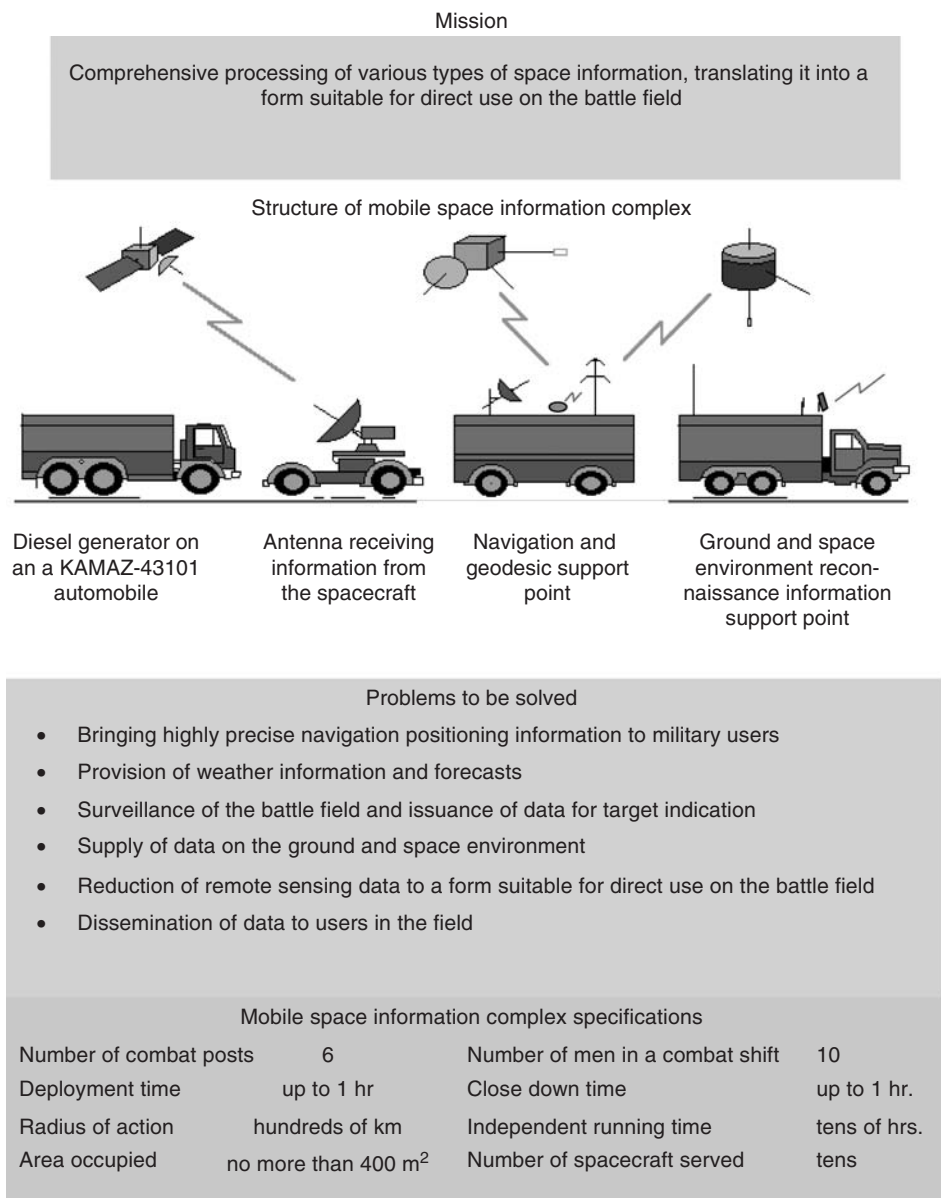


Figure 8. Mobile space information complex. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

integrating space technology components into systems and complexes of equipment, primarily high-precision (Fig. 9), and also the development of new technologies that can be used in the information war, rather than producing lethal effects (Fig. 10). Devices based on these technologies may be placed on spacecraft and will be capable of producing constant or periodic mass effects on selected

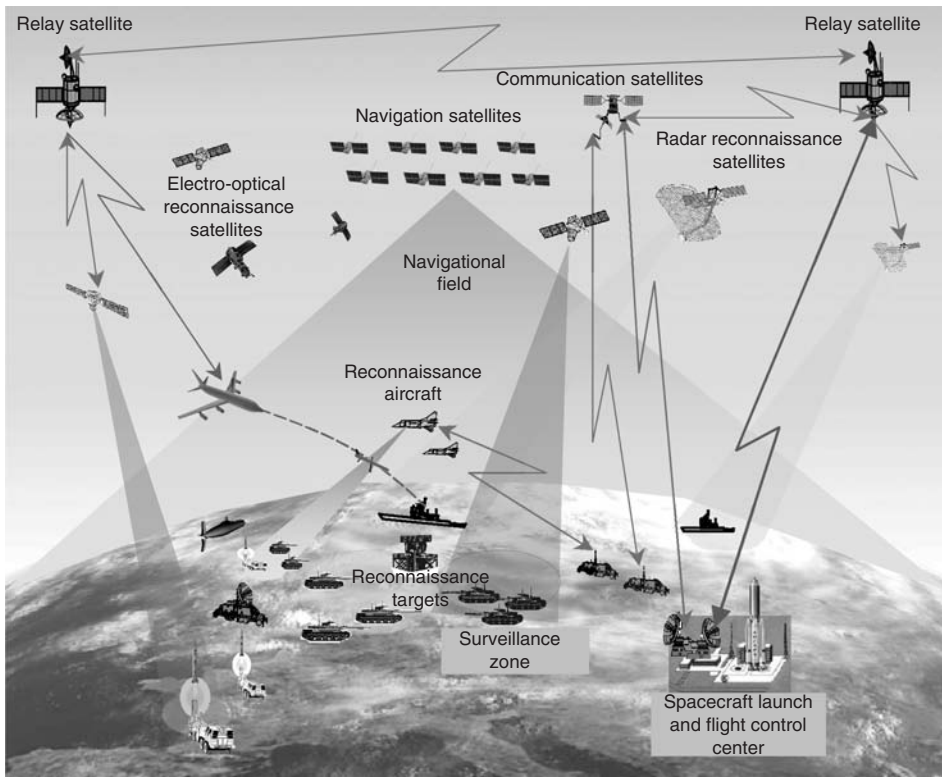


Figure 9. Integrated space, air, and ground-based reconnaissance and target designation system. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

regions with the goal of temporarily putting human forces out of commission, demoralizing the population, etc.

The potential for solving such problems from space will lead to a qualitative change in the form and methods of combat operations and in the organization of combat overall. The major goal of an “information war” is disintegration and disruption of the cohesion of command and control of enemy groupings, turning them into isolated, disoriented, and unmanageable elements that can subsequently be rendered harmless through physical destruction. The broad spectrum of issues related to the military uses of space shows that uncontrolled operations in this area could lead to catastrophic consequences, which is inevitably a cause of concern to the world community and thus will lead to the creation of international legal standards regulating military operations in space.

The rampant development of space systems and the enhancement of their role in maintaining the combat-readiness of modern armed forces has led to the irreversible militarization of near-Earth space in the interests of solving problems of reconnaissance, communications, navigation, and meteorological support, and missile defense. It is widely believed that such functions of space technology have a stabilizing effect and comply with Article 51 of the UN Charter, supporting nations’ right to self-defense.

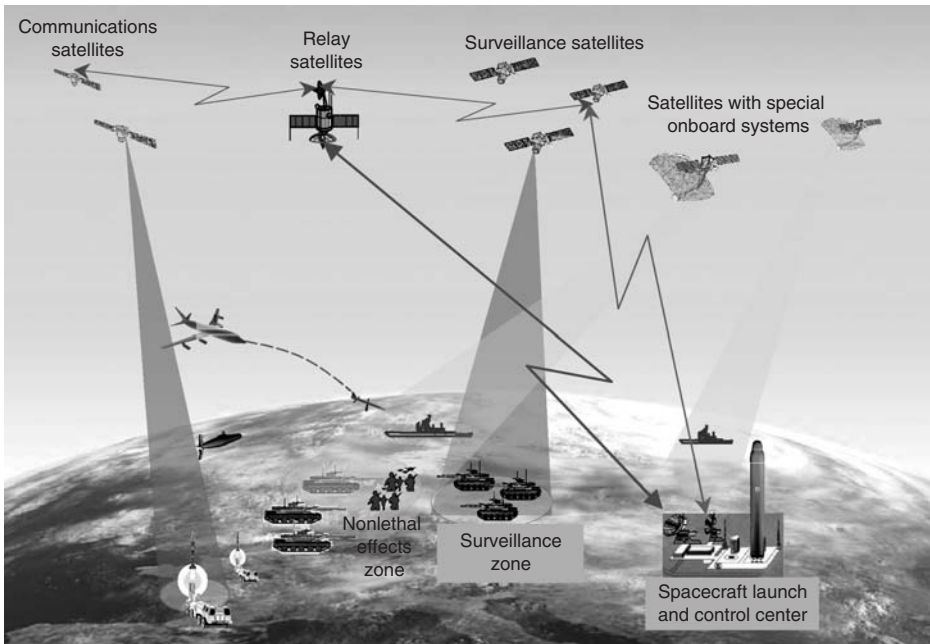


Figure 10. Satellites and nonlethal special onboard systems. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The military aspect of modern cosmonautics is thus manifest in the implementation of military space operations, a term which is generally taken to mean any operation associated with work to study and use space technology to ensure and maintain defense and security. Military space operations may include applied military research and development, space-based communications, reconnaissance, strategic defense, navigational, topogeodesic and meteorological support from space; the solution of particular problems in antisatellite and antimissile defenses, and also other forms of activity implemented in the interests of military and semimilitary agencies.

Contemporary military space operations, strictly speaking, are very weakly regulated by international law for the following reasons. First, because of the relatively recent and limited use of space for strictly military purposes, the need to develop exhaustive legal standards for military space operations has still not been universally acknowledged, and any real likelihood of developing them in the near future remains extremely low. Second, the term “military space operations” has not yet been either officially defined or incorporated in international law (the definition cited is merely one of the many that are in practical current use.) The multiple interpretations of this term result from the fact that such operations, especially those implemented during peacetime, may be seen as furthering the interests of humanity (from the standpoint of maintaining strategic stability and international security). At the same time, its destabilizing functions are capable of involving the international community in a qualitatively new round of the arms race and all of its negative military-strategic, political, technical, economic,

and other consequences. Third, international space law has no unambiguous and exhaustive interpretation of the concept “operations for peaceful purposes.” A more widespread concept, but one that has still not been officially adopted concept is “dual use technology” (intended for solving of purely peaceful, as well as strictly military applied problems in and from space). Thus, the current non-codified regime of military space activity is built on one of the basic principles of international law—“anything not specifically forbidden is permitted.” But despite this fact, military space operations could, with a greater or lesser degree of arbitrariness, equally be interpreted as forbidden, permitted, or not specified by international law.

One of the directions of military space operations, which may require legal regulation, involves adverse man-made effects on the space environment. Such effects may result from traditional (increase in the scale of space operations) as well as from military strategic factors (the attempt to anticipate and prevent implementation of aggressive actions in near-Earth space). Thus, as humanity continues to conquer space and new space technologies are developed, the applied military role of space in ensuring national security of states will certainly increase.

BIBLIOGRAPHY

1. Kiselev, A.I., A.A. Medvedev, and V.A. Menshikov. *Cosmonautics on the Threshold of the Millennium. Conclusions and Prospects*. Mashinostroyeniye, Moscow, 2001.
2. Menshikov, V.A. Russia's military space policy in the twenty-first century. *Voyennaya Mysl*, 6 (2000).
3. Space guarding the homeland. *Proceedings of the second lecture series in honor of M.K. Tikhonravov*, V.A. Menshikov (ed.). Kosmos, Moscow, 1999.
4. Menshikov, V.A., and S.V. Pavlov. NATO space technology in support of military operations in the Balkans. *Aerokosmicheskaya kuryer*, 4 (1999).
5. Menshikov, V.A. Concepts of development of military cosmonautic under conditions of the Armed Forces reform. *Dvoynnye tekhnologii*, [Dual Technologies]. 1 (1999).
6. Menshikov, V.A., I.N. Golovanev, and S.V. Pavlov. The soldier of the future. *Armeyskiy sbornik*, 2 (1997).
7. Menshikov, V.A. Military aspects of cosmonautics. *Proc. Cong. Russ. Fed. Cosmonautics*, 1997.

VALERIY A. MENSHIKOV
Khrunichev Space Center
Moscow, Russia

MOON

Introduction

The Moon, the natural satellite of Earth, has positively affected our development in many profound ways. Its orbital presence helps stabilize Earth's axial precession and thus, prevents the alternating extremes of climate that some planets,

such as Mars, experience. This equitable climate makes Earth a habitable world. The ocean tides induced by the Moon probably permitted vertebrate life to emerge from the sea more than 300 million years ago, leading to the development of land fauna and ultimately, to ourselves. Thus in a very real sense, we are here because the Moon exists.

Now, the Moon is exerting its beneficial influence on us once again by its existence. The Moon is the first milepost in humanity's movement into the solar system. We must pass by the Moon's orbit to go anywhere else in space. Thus, nature has provided us with a natural "way station," a place to learn how to live and work in space, refuel, and refresh our spacecraft. In addition to these benefits, the Moon also happens to be a rather interesting place. Its surface contains a record of the important events that occurred in the early history of Earth. Moreover, it is an excellent platform to observe the Universe around us. In all of these ways, the Moon is an important part of our movement into space.

Basic Properties and Motions

The Moon is quite large in relation to the planet it orbits, about 1% of the mass of Earth and about one-fifth its radius (Table 1). In surface area, the Moon is roughly the size of the continent of Africa, about 38 million square kilometers. Because the tenuous atmosphere of its surface is a near-perfect vacuum, no weather affects its terrain, and the sky is perpetually black. Stars are visible from the surface during daytime but are difficult to see because the glare reflected from the surface dilates your pupils. At high noon, the surface temperature can be more than 100°C and at midnight, as low as –150°C. The lunar day

Table 1. **Basic Data about the Moon**

Mass	7.35×10^{22} kg (1% of mass of Earth)
Radius	1738 km (27% of radius of Earth)
Surface area	3.79×10^7 km ² (7% of area of Earth)
Density	3340 kg/m ³ (3.34 g/cm ³)
Gravity	1.62 m/s ² (0.17 of Earth)
Escape velocity	2.38 km/s
Orbital velocity	1.68 km/s
Inclination of spin axis (to Sun)	1.6°
Inclination of orbital plane (to Sun)	5.9°
Distance from Earth	
Closest	356,410 km
Farthest	406,697 km
Orbital eccentricity	0.055
Albedo (fraction light reflected) average	0.07–0.24 (average terrae: 0.11–0.18; maria: 0.07–0.10)
Rotational period (noon-to-noon; average)	29.531 Earth days (709 hours)
Average surface temperature	107°C day; –153°C night
Surface temperature in polar areas	–30° to –50°C in light; –230°C in shadows

(the time it takes to rotate once on its spin axis) is about 29 Earth days or 708 hours, and daylight hours on the Moon (sunrise to sunset) last almost 2 weeks. The Moon is famous for its low gravity, about one-sixth of Earth's. Thus, an astronaut who weighs 200 pounds on Earth weighs only 34 pounds on the Moon.

The Moon moves in an elliptical path around Earth and completes its circuit once every 29 days. This time is equal to the amount of time it takes for the Moon to rotate once on its axis (the lunar day). In consequence, the Moon shows the same hemisphere (called the *near side*) to Earth at all times. Conversely, one hemisphere is forever turned away from us (the *far side*) (Fig. 1). Before the space age, the far side of the Moon was completely unknown territory, not revealed to human gaze until its face was first photographed by the Soviet spacecraft Luna 3 in 1959.

The elliptical orbit of the Moon results in a variable distance between Earth and Moon. At perigee (when the Moon is closest to Earth), the Moon is a mere 356,410 km away; at apogee (the farthest position), it is 406,697 km away. This is different enough so that the apparent size of the Moon in the sky varies; its average apparent size is a little smaller than that of a dime held at arm's length. In works of art, a huge lunar disk looming above the horizon is often depicted, but such an appearance is an illusion. A Moon near the horizon can be compared in size with distant objects on the horizon, such as trees, making it seem large, and a Moon near zenith (overhead) cannot be compared easily with earthly objects, and hence, seems smaller.

The plane of the Moon's orbit lies neither in the equatorial plane of Earth nor in ecliptic plane, in which nearly all the planets orbit the Sun (Fig. 2). This relation poses some constraints on models of lunar origin. The spin axis of the

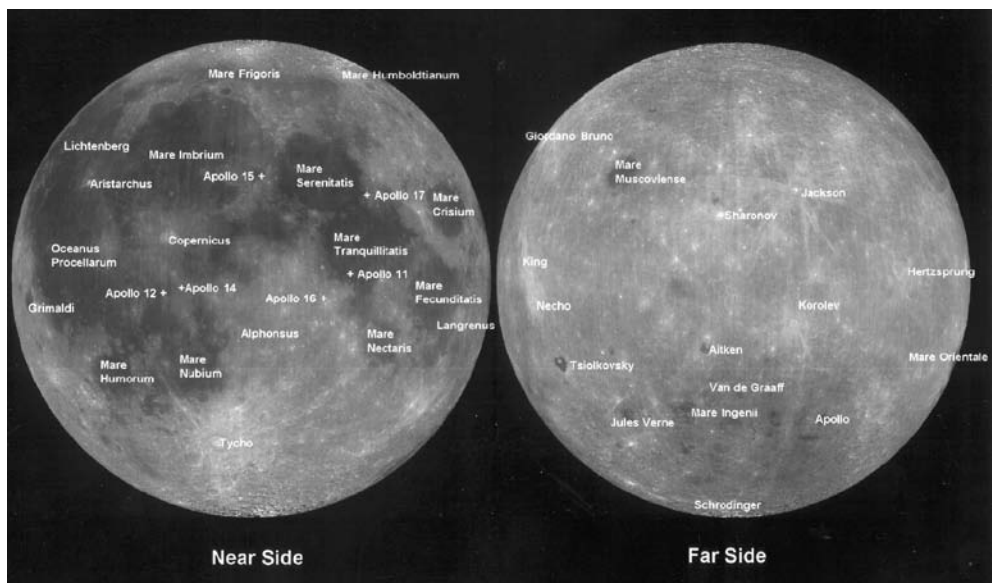


Figure 1. Index map of the Moon (Clementine albedo image), showing the location of some prominent lunar features. Apollo landing sites are shown by crosses.

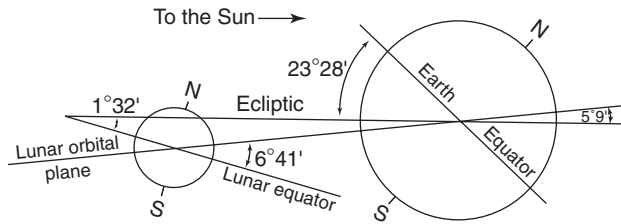


Figure 2. Orbital planes and spin axes of Earth and Moon. Although Earth's axis is tilted 23° from the ecliptic, the Moon's is nearly perpendicular to it, resulting in grazing solar illumination near the poles.

Moon is nearly perpendicular to the ecliptic plane; it has an inclination of about 1.5° from the vertical. This simple fact has some really significant consequences. Because its spin axis is vertical, the Moon experiences no “seasons”, as does Earth, whose inclination is about 24° . So, as the Moon rotates on its axis, an observer at the pole would see the Sun hovering close to the horizon. A large peak near the pole might be in permanent sunlight, and a crater floor could exist in permanent shadow. In fact, we now know that such areas exist, particularly near the South Pole. The existence of such regions has important implications for a return to the Moon.

As the Moon circles Earth, they occasionally block the Sun for each other, causing *eclipses*. A *solar eclipse* occurs when the Moon is between the Sun and Earth and can occur only at new Moon (the dayside of the Moon is facing the Sun). Because of the variable distance between Earth and the Moon, its inclined orbital plane, and the smaller size of the Moon, solar eclipses are quite rare (years may pass between total solar eclipses), so their occurrence is always subject to much hoopla. A *lunar eclipse*, in contrast, occurs when the Earth is between the Moon and the Sun. These events happen much more frequently, because Earth's shadow has a much larger cross-sectional area than the Moon's shadow. Lunar eclipses can occur only during a full Moon (or new Earth). As the shadow of Earth slowly covers the full Moon, it takes on a dull red glow, caused by the bending of some sunlight that illuminates the Moon through the thick atmosphere of Earth.

The Moon is gradually receding from Earth. Early in planetary history, Earth was spinning much faster, and the Moon orbited much closer than it does now. Over time, energy has been transferred from Earth to the Moon, causing the spin rate of Earth to decline and the Moon to speed up in its orbit, thus moving farther away (the current rate of recession is about 4 cm/year). Such recession will continue; some day, the Moon will be too far away to create a total solar eclipse! Fortunately for lovers of cosmic spectacles, this will not happen for at least another few million years.

As the Moon orbits Earth, we can peek around its edges because of a phenomenon known as *libration*. Libration in latitude is caused by the 7° inclination of the plane of the Moon's orbit to Earth's equator. This inclination allows us to “look over the edge” of the Moon as it moves slightly above or slightly below the equatorial plane. Libration in longitude is caused by the Moon's elliptical orbit, which permits Earth viewers to look around its leading or trailing edge. A small

libration is also caused by parallax, which is the effect that allows you to see more by moving side to side, in this case by the diameter of Earth. All told, these libration effects permit us to see slightly more than a single hemisphere, and over time, we can see about 59% of the lunar surface.

The gravitational influence of Earth and Moon upon each other is considerable. Because of the gravitational tug of the Moon and Sun, the Earth experiences *tides*, which are bulges in the radius of Earth induced by gravitational attraction. Tides, it is often thought, are associated with the oceans, but solid Earth also undergoes an up and down motion caused by tides. Because Earth attracts the Moon just as much in reverse, the Moon also experiences a tidal bulge, one that mirrors the tidal effects on Earth. The raising and lowering of solid body tides on Earth and Moon causes friction inside the two planets, and this source of heat is called tidal dissipation. Such an energy source for planetary heat may have been very important early in the history of the solar system, when Moon and Earth were closer together, but it is currently only a minor source of heat.

Origin of the Moon and its Structure

In their surveys of the solar system, astronomers have discovered dozens of satellites around other planets. Yet, of the four inner planets, only Earth and Mars have moons (and Mars' are probably captured asteroids). Ours is remarkably large as satellites go, particularly compared to the modest size of Earth itself. The creation of the Moon was thus an unusual event in terms of general planetary evolution, and our knowledge of the solar system—however detailed—would be profoundly incomplete without determining how our enigmatic satellite came to exist.

Traditionally, scientists have investigated three models of lunar origin. In the simplest hypothesis, termed *co-accretion*, Earth and Moon formed together from gas and dust in the primordial solar nebula and have existed as a pair from the outset. A second concept, called the *capture* scenario, envisions the Moon as a maverick world that strayed too near the Earth and became trapped in orbit—either intact or as ripped-apart fragments—due to our planet's strong gravity. According to the third model, termed *fission*, the Earth initially had no satellite but somehow began to spin so fast that a large fraction of its mass tore away to create the Moon (1).

It was hoped that our astronauts would return with results that would allow us to choose decisively from among these three models. Study of the Apollo samples has provided some constraints on the true lunar origin, but none of these models has proven completely satisfactory. First, the Moon's bulk composition appears to be similar, but not identical, to the composition of Earth's upper mantle. Both are dominated by the iron- and magnesium-rich silicates pyroxene and olivine. But one important distinction is that, unlike Earth, the Moon generally lacks volatile elements. Another involves the relative dearth in lunar material of what are termed siderophile ("metal-loving") elements such as cobalt and nickel, which tend to occur in mineral assemblages that contain metallic iron (Fig. 3).

A second key constraint comes from oxygen's three natural isotopes: ^{16}O , ^{17}O , and ^{18}O . Ratios of these isotopes are identical in lunar and terrestrial

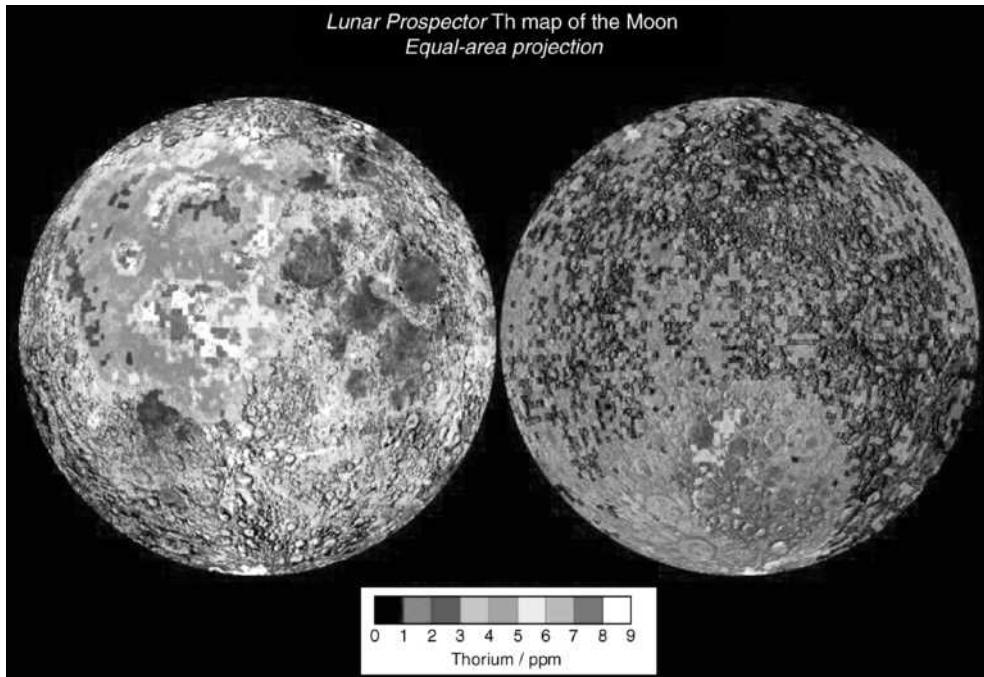


Figure 3. Maps of the epithermal neutron flux near the South Pole of the Moon from the Lunar Prospector. Low epithermal flux indicates the presence of hydrogen. This map shows that the highest hydrogen concentrations are associated with areas of permanent darkness near the South Pole. In conjunction with positive radar evidence from Clementine, this indicates that water ice exists near the South Pole of the Moon. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

materials, which suggests strongly that the Moon and Earth originated in the same part of the solar system. These same ratios are different in meteorites such as eucrites (asteroidal basalts), the so-called SNC group (possible Martian igneous rocks), and various subgroups of ordinary chondrites.

Beyond geochemical evidence, several physical properties of the Earth–Moon system provide important clues in determining lunar origin. For example, the pair possesses a great deal of angular momentum. Also, the Moon’s orbit does not lie within the plane of Earth’s equator or of its orbit (the ecliptic plane). Finally, the Moon is gradually receding from Earth at roughly 3 cm per year—a curious effect caused by the gravitational coupling of the Moon and our oceans. Tidal bulges raised in seawater do not lie directly along the Earth–Moon line but actually precede it, because Earth’s rotation drags them along for some distance before they can adjust to the Moon’s changing location in the sky. This misalignment causes Earth’s rotation to decelerate slightly; the Moon in turn is pulled forward in its orbit, speeds up, and inches farther away. Unfortunately, we cannot determine the original Earth–Moon distance because the orbital recession going on now cannot be extrapolated back to the lunar origin.

A new idea for the birth of the Moon has gained popularity recently and even something of a consensus, although all the attendant problems that it poses

have yet to be resolved. This idea is that a giant object, possibly a planet-sized body as big as Mars, hit Earth around 4.6 billion years ago (Fig. 4). It may have struck off-center, thereby increasing Earth's rotational rate. A mixture of terrestrial and impactor material would have been thrown into Earth orbit and later coalesced to form the Moon.

Because this material would have jetted into space in a predominantly vaporized state, the *giant-impact* hypothesis could explain both the Moon's dearth of volatile elements and its possible slight enrichment in refractory elements (those that remain solid at high temperature). To account as well for the Moon's depletion in metallic iron and siderophile elements relative to Earth, theorists must assume that the incoming object had already differentiated into a core and mantle. Their calculations show that at least half to nearly all of the lunar mass was derived from the outer layers of the colliding body. So to create a proper Moon depleted in iron and siderophiles, these elements would have to be concentrated in the impactor's core, which became incorporated into the Earth shortly after the initial collision. [See article Appollo 17 and the Moon by Harrison Schmitt elsewhere in this Encyclopedia.]

The giant-impact hypothesis appears to explain, or allow for, several fundamental relations—not just bulk composition, but also the orientation and evolution of the lunar orbit. It also makes the uniqueness of the Earth–Moon system seem more plausible, that is, impacts of this magnitude might have occurred only rarely, rather than as a requirement for planetary formation. Part of the reason for this model's current popularity is doubtless because we know too little to rule it out: key factors such as the impactor's composition, the collision geometry, and the Moon's initial orbit are all underdetermined.

Scientists realize that the advent of the giant-impact hypothesis has not “solved” the problem of lunar origin. For example, the close genetic relation of Earth and Moon (inferred from the oxygen-isotope ratios) is not an obvious consequence of a giant impact, especially if most of the lunar mass derived from

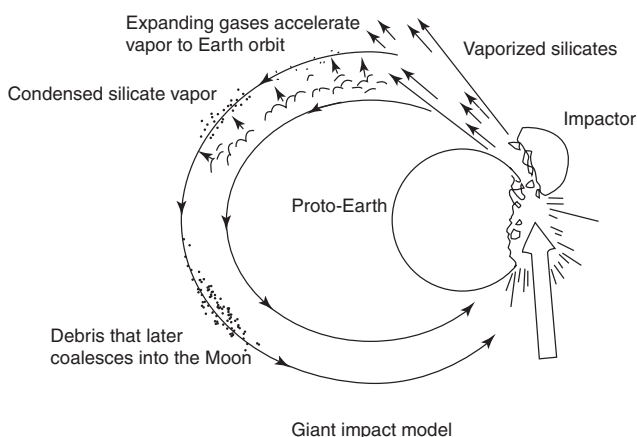


Figure 4. Giant impact model of lunar origin. In this hypothesis, a Mars-sized planet hit the Earth at grazing incidence, throwing vaporized mantle into Earth orbit. This material later coalesced to form the Moon (after Reference 8).

the projectile. Consequently, research into the effects of such cataclysmic impacts in early planetary history continues at a brisk pace. But this model of lunar origin appears to explain the most salient features of the Moon and has a minimum amount of special pleading.

Scientific Results from Apollo

From 1969 to 1972, six Apollo expeditions set down on the Moon, allowing a dozen American astronauts to explore the lunar landscape and return with pieces of its surface (Fig. 1). The initial landing sites were chosen primarily on the basis of safety. Apollo 11 landed on the smooth plains of Mare Tranquillitatis, Apollo 12 on a mare site near the east edge of the vast Oceanus Procellarum. These first missions confirmed the volcanic nature of the maria and established their antiquity (older than 3 billion years). Later missions visited sites of increasing geologic complexity. Apollo 14 landed in highland terrain near the crater Fra Mauro, an area, it was thought, is covered by debris thrown out by the impact that formed the Imbrium basin. Apollo 15 was the first mission to employ a roving vehicle and the first sent to a site that contains both mare and highland units (the Hadley–Apennine region). Apollo 16 landed on a highland site near the rim of the Nectaris basin. The final lunar mission in the series, Apollo 17, was sent to a combination mare–highland site on the east edge of the Serenitatis basin (2).

The Soviet Union has acquired a small but important set of lunar samples of its own, thanks to three automated spacecraft that landed near the eastern limb of the Moon’s near side. Luna 16 visited Mare Fecunditatis in 1970, and Luna 24 went to Mare Crisium in 1976. A third site, in the highlands surrounding the Crisium basin, was visited by Luna 20 in 1972.

Altogether, these nine missions returned 382 kg of rocks and soil, the “ground truth” that provides most of our detailed knowledge of the Moon. Though the most exhilarating discoveries came from studies completed years ago, today scientists around the world continue to examine these samples, establish their geologic contexts, and make inferences about the regional events that shaped their histories. What we have learned about the Moon’s three major surface materials—maria, terrae, and the soil-like regolith that covers both—is summarized in the following paragraphs.

Regolith. During the Moon’s history, micrometeorite bombardment thoroughly pulverized the surface rocks into a fine-grained, chaotic mass of material called the regolith (also informally called “lunar soil,” though it contains no organic matter). The regolith consists of single mineral grains, rock fragments, and combinations of these that have been cemented by impact-generated glass. Because the Moon has no atmosphere, its soil is directly exposed to the high-speed solar wind, gases flowing out from the Sun that become implanted directly onto small surface grains. The regolith’s thickness depends on the age of the bedrock that underlies it and thus how long the surface has been exposed to meteoritic bombardment; the regolith in the maria is 2–8 meters thick, whereas in highland regions its thickness may exceed 15 meters.

The composition of the regolith closely resembles that of the local underlying bedrock. Some exotic components are always present, perhaps having

arrived as debris flung from a large distant impact. But this is the exception rather than the rule. The contacts between mare and highland units appear sharp from lunar orbit, which suggests that relatively little material has been transported laterally. Thus, although mare regoliths may contain numerous terrae fragments, in general these derive not from far-away highland plateaus but are instead crustal material excavated locally from beneath the mare deposits.

Impacts energetic enough to form meter-size craters in the lunar regolith sometimes compact and weld the loose soil into a type of rock called *regolith breccia*. Once fused into a coherent mass, a regolith breccia no longer undergoes the fine-scale mixing and “gardening” taking place in the unconsolidated soil around it. Thus, regolith breccias are “fossilized soils” that retain their ancient composition and also the chemical and isotopic properties of the solar wind from the era in which they formed.

Maria. Thanks to our lunar samples, there is no longer any doubt that the maria are volcanic. Mare rocks are *basalts* that have a fine-grained or even glassy crystalline structure (indicating that they cooled rapidly) and are rich in iron and magnesium. Basalts are a widespread volcanic rock on Earth, consisting mostly of the common silicates pyroxene and plagioclase, numerous accessory minerals, and sometimes olivine (an iron-magnesium silicate). But lunar basalts display some interesting departures from this basic formulation. For example, they are completely devoid of water—or any form of hydrated mineral—and contain few volatile elements in general. Basalts from Mare Tranquillitatis and Mare Serenitatis are remarkably abundant in titanium; sometimes they contain roughly 10 times more than is typically found in their terrestrial counterparts.

Mare basalts originated hundreds of kilometers deep within the Moon in the total absence of water and the near absence of free oxygen. There, the heat from decaying radioactive isotopes created zones of partially molten rock that ultimately forced its way to the surface. The occurrence of mare outpourings within impact basins is no chance coincidence, for the crust beneath these basins must have been fractured to great depth by the cataclysmic impacts that formed them. Much later, molten magmas rose to the surface through these fractures and erupted onto the basin floors.

Although they may appear otherwise, the maria average only a few hundred meters in thickness. These volcanic veneers tend to be thinner near the rims that confine them and thicker over the basins’ centers (as much as 2–4 km in some places). What the maria may lack in thickness they make up for in sheer mass, which frequently is great enough to deform the crust underneath them. This has stretched the outer edges of the maria (creating fault-like depressions called grabens) and compressed their interiors (creating raised “wrinkle” ridges) (Fig. 5).

Basalts returned from the mare plains range in age from 3.8–3.1 billion years, a substantial interval of time. But small fragments of mare basalt found in highland breccias solidified even earlier—as long ago as 4.3 billion years. We do not have samples of the youngest mare basalts on the Moon, but stratigraphic evidence from high-resolution photographs suggests that some mare flows actually embay (and therefore postdate) young, rayed craters and thus may be no older than 1 billion years.

A variety of volcanic glasses—distinct from the ubiquitous, impact-generated glass beads in the regolith—were found in the soils at virtually all of the

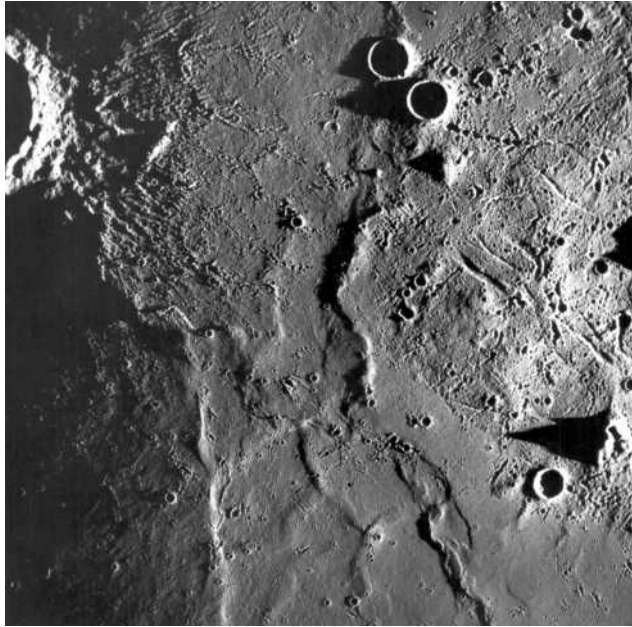


Figure 5. Area in eastern Mare Imbrium, showing wrinkle ridges (left), representing compressional tectonics, and grabens (right), representing extensional tectonics. These two landforms are the typical tectonic features of the Moon.

Apollo landing sites. They even were scattered about the terrae sites, far from the nearest mare. Some of these volcanic materials are similar in chemical composition, but not identical, to mare basalts and were apparently formed at roughly the same time.

One such sample, tiny beads of orange glass, came from the Apollo 17 site. They are akin to the small airborne droplets that accompany volcanic “fire fountains” on Earth, like those in Hawaii. The force of the eruption throws bits of lava high into the air, and they solidify into tiny spherules before hitting the ground. The Moon’s volcanic glass beads have had a similar origin. The orange ones from the Apollo 17 site get their color from high titanium content, more than 9%, and some of them are coated with amorphous mounds of volatile elements like zinc, lead, sulfur, and chlorine (see Appollo 17 and the Moon in this volume).

Terrae. One could easily imagine that the lunar highlands contain outcrops of the original lunar crust—much as we find in Earth’s continents. But what really awaited the astronauts was a landscape so totally pulverized that no traces of the original outer crust survived intact. Instead, most of the stones collected from the terrae were breccias, usually containing fragments from a wide variety of rock types that have been fused together by impact processes. Most of these consist of still older breccia fragments that attest to a long and protracted bombardment history.

The highland samples also include several fine-grained crystalline rocks that have a wide range of compositions. They are not breccias, but they were created during an impact. In these cases the shock and pressure were so overwhelming that the “target” melted completely and created in effect entirely new

rocks from whatever ended up in the molten mass. The impactors become part of this mixture, and these impact-melt rocks contain distinct elemental signatures of meteoritic material.

Based on the samples in hand, virtually all of the highlands' breccias and impact melts formed between about 4.0 and 3.8 billion years ago (Fig. 6). The relative brevity of this interval surprised researchers—why were all the highland rocks so similar in age? Perhaps the rate of meteoritic bombardment on the Moon increased dramatically during that time. Alternatively, the narrow age range may merely mark the conclusion of an intense and continuous bombardment that began 4.6 billion years ago, the estimated time of lunar origin. To resolve the enigma, we must return to the Moon and sample its surface at carefully selected geologic sites.

A substantial number of small, whitish rock fragments found in the mare soils returned by Apollo 11 and 12 astronauts had compositions totally unlike that of basalts and virtually unmatched on Earth. They consisted almost entirely of plagioclase feldspar, a silicate rich in calcium and aluminum but depleted in heavier metals such as iron. A few prescient researchers postulated that these rocks came from the lunar highlands. The last four Apollo missions, sent to highland landing sites, confirmed that plagioclase feldspar dominates the lunar crust. The resulting implication was broad and profound: at some point in the distant past much of the Moon's exterior—and perhaps its entire globe—had been molten.

The detailed nature of this waterless “magma ocean” is only dimly perceived at present; for example, the lunar surface may not have been completely molten everywhere. But the consequences seem clear. In a deep, slowly cooling layer of lunar magma, crystals of low-density plagioclase feldspar would have

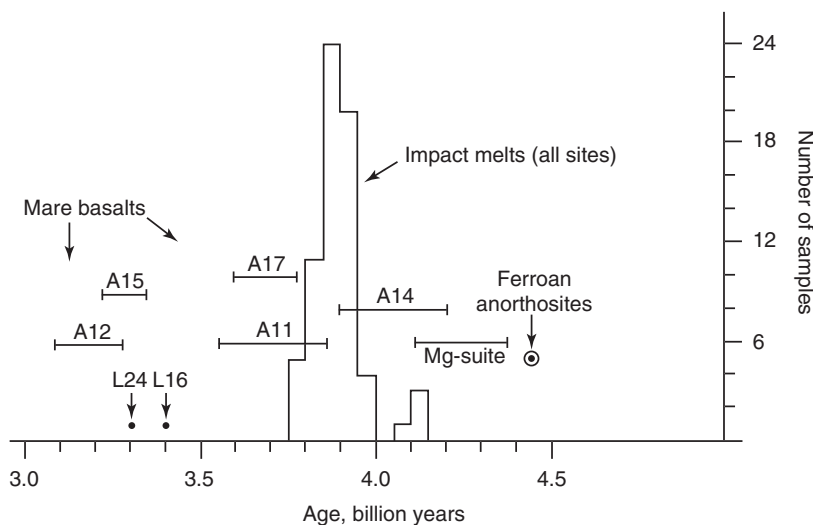


Figure 6. Histogram of lunar rock ages. Note that most lunar rocks are extremely ancient and date from 3–4.5 billion years (the solar system is 4.6 billion years). Virtually all of the impact melts from the lunar highlands date from the narrow time interval between 3.8 and 4.0 billion years ago, a time known as the lunar “cataclysm.”

risen upward after forming, and higher density minerals would have accumulated at lower levels. This segregation process, termed *differentiation*, left the young Moon with a crust that was, in effect, a low-density rock “froth” tens of kilometers thick that consists mostly of plagioclase feldspar. At the same time, denser minerals (particularly olivine and pyroxene) became concentrated in the mantle below—the future source region of mare basalts.

It is unclear to what depth the magma ocean extended, but the coatings of volatile elements discovered on some mare glasses provide an important clue. If the Moon’s exterior really was once molten, the most volatile components in the melt would have vaporized and escaped into space. But the volatile-coated glasses sprayed onto the lunar surface long after the magma ocean solidified. If the glasses’ compositions did not change in their upward migration from the lunar interior, they imply that volatile-rich pockets remained (and perhaps still exist) in the upper mantle. The implication, therefore, is that the magma ocean was at most only a few hundred kilometers deep (3).

The highland samples returned by the last four Apollo crews provided other surprises. Unlike glasses and basalts, which quench quickly after erupting onto the surface, some of the clasts in the highland breccias contained large, well-formed crystals, indicating that they had cooled and solidified slowly, deep inside the Moon. These igneous rocks sometimes occur as discrete specimens. At least two distinct magmas were involved in their formation. Rocks composed almost completely of plagioclase feldspar, that have just a hint of iron-rich silicates are called *ferroan anorthosites*. They are widespread in the highlands. Absolute dating of the anorthosites has proved difficult, but it appears that they are extremely ancient, having crystallized very soon after the Moon formed (4.6 to 4.5 billion years ago).

The highlands’ other dominant rock type is also abundant in plagioclase feldspar, but it contains substantial amounts of olivine and a variety of pyroxene low in calcium. This second class of rocks is collectively termed the *Mg-suite*, so called because they contain considerable magnesium (Mg). These rocks appear to have undergone the same intense impact processing as the anorthosites, and their crystallization ages vary widely—from about 4.3 billion years to almost the age of the Moon.

The anorthosite and Mg-suite rocks could not have crystallized from the same “parent” magma, so at least two (and probably more) deep-seated sources contributed to the formation of the early lunar crust. Conceivably, both magmas might have existed simultaneously during the first 300 million years of lunar history. This would contradict our notion of the Moon as a geologically simple world and greatly complicate our picture of the formation and early evolution of its crust.

During early study of the Apollo samples, an unusual chemical component was identified that is enriched in incompatible trace elements—those that do not fit well into the atomic structures of the common lunar minerals plagioclase, pyroxene, and olivine as molten rock cools and crystallizes. This element group includes potassium (K), rare-earth elements (REE) such as samarium, and phosphorus (P); geochemists refer to this element combination as *KREEP*. It is a component of many highland soils, breccias, and impact melts, yet the trace-element abundances remain remarkably constant wherever it is found. Moreover, its estimated age is consistently 4.35 billion years. These characteristics

have led to the consensus that KREEP represents the final product of the crystallization of a global magma system that solidified aeons ago.

But the evidence of chemically distinct, widespread volcanic rocks in the highlands—KREEP-rich or otherwise—remains tenuous. Some highland rocks are compositionally similar to mare basalts yet exhibit KREEP's trace-element concentrations. For example, the Apollo 15 astronauts returned with true basalts that probably derive from the nearby Apennine Bench Formation (Fig. 5), a large volcanic outflow situated along the Imbrium basin's rim. These "KREEP basalts" have a well-determined age of 3.85 billion years, so the Imbrium impact must have occurred before this date and probably just before the Apennine Bench Formation extruded onto the surface. Thus, although the extent and importance of highland volcanism remains unknown, it apparently took place early in lunar history and contributed at least some of the KREEP component observed in highland breccias and impact melts (4).

Water on the Moon

An abundant supply of water on the Moon would make establishing a self-sustaining lunar colony much more feasible and less expensive than presently thought. Study of lunar samples revealed that the interior of the Moon is essentially devoid of water, so no underground supplies could be used by lunar inhabitants. However, the lunar surface is bombarded with water-rich objects such as comets, and scientists have suspected that some of the water in these objects could migrate to permanently dark areas at the lunar poles and perhaps accumulate to usable quantities.

Water is constantly being added to the Moon from impacting comets and water-rich asteroids. Where would this water end up? Most of it would be split by sunlight into its constituent atoms of hydrogen and oxygen and lost into space, but some would migrate by literally hopping along to places where it is very cold. It was postulated that the polar regions might have areas that are permanently shadowed, hence permanently cold. The water might accumulate there.

The Moon's axis of rotation is nearly perpendicular to the plane of its orbit around the Sun (Fig. 2). Although the plane of the Moon's orbit about Earth is inclined about 5° , its equator is inclined about 6.5° , resulting in a 1.5° inclination of the Moon's spin axis to its orbital plane around the Sun. This means that the Sun always appears close to the horizon at the poles of the Moon.

It has been calculated that temperatures in these permanently dark areas may be as cold as 40 to 50 K (-230° to -220°C), only a few tens of degrees above absolute zero. Moreover, these "cold traps" have existed on the Moon for at least the last 3–4 billion years—plenty of time to accumulate water from impacting comets.

To determine whether there is water on the Moon, we had to await the results of polar-orbiting, global mappers. Two missions, Clementine and Lunar Prospector, sent to the Moon in the 1990s, looked for evidence of water at the poles.

Results from Post-Apollo Robotic Missions

Clementine. Clementine was a mission designed to test the spaceworthiness of a variety of advanced sensors for use on military surveillance satellites and, at

the same time, to gather useful scientific information on the composition and structure of the Moon and a near-Earth asteroid. Conducted jointly by the Ballistic Missile Defense Organization (BMDO, formerly the Strategic Defense Initiative Organization) of the U.S. Department of Defense and NASA, Clementine was sent for an extended stay in the vicinity of Earth's Moon on 25 January 1994 and arrived at the Moon on 20 February 1994. The spacecraft started systematic mapping on 26 February 1994, completed mapping on 22 April 1994, and left lunar orbit on 3 May 1994 (5).

During 71 days in lunar orbit, Clementine systematically mapped the 38 million square kilometers of the Moon in 11 colors in the visible and near-infrared parts of the spectrum. In addition, the spacecraft took tens of thousands of high-resolution and mid-infrared thermal images, mapped the topography of the Moon with a laser ranging experiment, improved our knowledge of the surface gravity field of the Moon through radio tracking, and carried a charged particle telescope to characterize the solar and magnetospheric energetic particle environment. We have had our first view of the global color of the Moon, identifying major compositional provinces; studied several complex regions, mapping their geology and composition in detail; measured the topography of large, ancient impact features, including the largest (2500 km diameter), deepest (more than 10 km) impact basin known in the solar system; and deciphered the gravitational structure of a young basin on the limb of the Moon, finding that a huge plug of the lunar mantle is uplifted below its surface.

The color of the Moon in the visible to near-infrared part of the spectrum is sensitive to variations in both the composition of surface material and the amount of time that material has been exposed to space. The Clementine filters were selected to characterize the broad lunar continuum and to sample parts of the spectrum that, it is known, contain absorption bands diagnostic of iron-bearing minerals. By combining information obtained through several filters, multispectral image data are used to map the distribution of rock and soil types on the Moon.

Preliminary studies of areas of already known geologic complexity allow us to identify and map the diversity within and between geologic units, which have both impact and volcanic origins. The Aristarchus Plateau is a rectangular, elevated crustal block about 200 km across, surrounded by the vast mare lava plains of Oceanus Procellarum. Clementine altimetry shows that the plateau is a tilted slab that slopes down to the northwest and rises more than 2 km above Oceanus Procellarum on its southeastern margin. The plateau was probably uplifted, tilted, and fractured by the Imbrium basin impact, which also deposited hummocky ejecta on the plateau surface.

The plateau has experienced intense volcanic activity, both effusive and explosive. It includes the densest concentration of lunar sinuous rilles, including the largest known, Vallis Schröteri, which is about 160 km long, up to 11 km wide, and 1 km deep. The rilles in this area begin at "cobra-head" craters, which are the apparent vents for low-viscosity lavas that formed the rilles. These and other volcanic craters may have been the vents for a "dark mantling" deposit that covers the plateau and nearby areas to the north and east. This dark mantling deposit probably consists primarily of iron-rich glass spheres (pyroclastics or cinders) and has a deep red color. Rather than forming cinder cones as on Earth,

the lower gravity and vacuum of the Moon allows the pyroclastics to travel much greater heights and distances and thus deposit an extensive regional blanket.

The Aristarchus impact occurred relatively recently in geologic time, after the Copernicus impact but before the Tycho impact. The 42-km diameter crater and its ejecta are especially interesting because of their location on the uplifted southeastern corner of the Aristarchus plateau. As a result, the crater ejecta reveal two different stratigraphic sequences: that of the plateau to the northwest and that of a portion of Oceanus Procellarum to the southeast. This asymmetry is apparent in the colors of the ejecta, which are reddish to the southeast, dominated by excavated mare lava, and bluish to the northwest, caused by the excavation of highlands materials in the plateau. The extent of the continuous ejecta blanket also appears asymmetrical: it extends about twice as far to the north and east than in other directions, approximately following the plateau margins. These ejecta lobes could be caused by an oblique impact from the southeast, or it may reflect the presence of the plateau during ejecta emplacement.

The Clementine multispectral data will enable us to reconstruct the three-dimensional composition and geologic history of this region. In this color-ratio composite, fresh highlands materials are blue, fresh mare materials are yellowish, and mature mare soils are purplish or reddish. The subsurface compositions, buried beneath a few meters or tens of meters of pyroclastics or Aristarchus ejecta, are revealed by craters that penetrated the surface layers and by steep slopes such as those along the walls of the rilles. From this mosaic, we have seen that the plateau is composed of a complex mixture of materials, but that the rilles formed primarily in lavas, except for the cobra-head crater of Vallis Schröteri that formed in highland materials.

The laser ranging data from Clementine has allowed us to see the large-scale topography (or relief) of the lunar surface on a nearly global basis (Fig. 7). A striking result from this experiment is confirmation of the existence of a population of very ancient, nearly obliterated impact basins, randomly distributed across the Moon. These basins had been postulated on the basis of obscure circular patterns on poor quality photographs; Clementine laser ranging has provided dramatic confirmation of their existence, including their surprising depth, ranging from 5–7 kilometers, even for the most degraded features. Gravity data obtained from radio tracking of Clementine indicates that these great holes in the Moon's crust are compensated for by plugs of dense rocks far below the surface; such dense rocks are probably caused by structural uplift of the mantle (the iron- and magnesium-rich layer below the low-density, aluminum-rich crust) beneath these impact basins. Finally, Clementine laser ranging data have shown us the dimensions of the largest confirmed basin on the Moon, the 2500-km diameter South Pole-Aitken basin (Fig. 7): this feature averages more than 12 kilometers deep that makes it the largest, deepest impact crater known in the solar system (6).

Although the Clementine spacecraft did not carry instruments designed to look for lunar ice, during the mission, we improvised an experiment that allowed us to address this question. Radio waves are reflected from planetary surfaces differently, depending on the compositional makeup of those surfaces. Specifically, radio waves are scattered in all directions by reflection from surfaces made

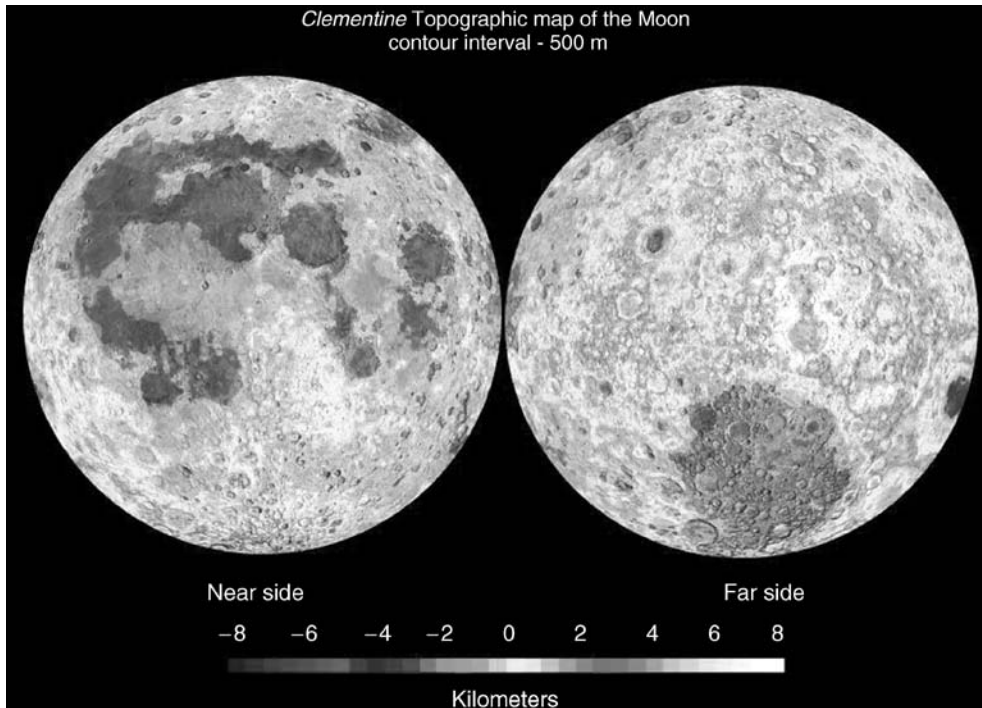


Figure 7. Global map of the topography of the Moon derived from Clementine laser altimetry and stereo photography (for the poles). The Moon shows a huge dynamic range of elevations, mostly caused by the presence of large, impact craters, called basins. Note the huge depression on the southern far side—this is the South Pole-Aitken basin, more than 2500 km in diameter and more than 12 km deep, the largest, deepest impact crater known in the solar system. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

up of ground-up rock (as are the terrestrial planets, which include most of the Moon, Mercury, Venus, Mars, and the asteroids), whereas radio waves are reflected more coherently from ice surfaces (the polar caps of Mercury and Mars and the icy surfaces of Jupiter's satellites Europa, Ganymede, and Callisto). When radio waves encounter ice, it is partly absorbed and reflected multiple times by internal flaws in the ice, then reflected back out into space. A consequence of multiple reflections is that some of the radio reflections come back in the same sense as they were transmitted (think of the reflection of light from two mirrors—reflection from a single mirror makes text unreadable, but double reflection makes the text “normal” again.) Thus, ice reflects back partly in the same sense as incident waves (5).

Analysis of the data returned from this radio-wave experiment performed in 1994 while the Clementine spacecraft was orbiting the Moon reveals that deposits of ice exist in permanently dark regions near the South Pole of the Moon. Initial estimates suggest that ice exists at both poles and has the volume of a small lake more than 1 billion cubic meters ($\sim 1 \text{ km}^3$). This amount of water would be equivalent to the fuel (hydrogen and oxygen) used for more than 100,000 launches of the Space Shuttle.

Lunar Prospector. The Lunar Prospector (LP), the first of NASA's new, cheaper, "Discovery"-class missions, was launched to the Moon on 6 January 1998. Prospector orbited the Moon in a 100-km altitude, polar orbit for more than 18 months. It carried a variety of instruments that, in many ways, complemented the instruments of the earlier Clementine mission. In part, Lunar Prospector's stated objectives were to map the resources of the Moon, including assessments of polar volatiles and basic elemental composition. In addition, it would map the gravity and magnetic fields of the Moon during its 1-year nominal mission.

The spacecraft was a spin-stabilized microsat, about 100 kg in dry mass. It carried a gamma-ray spectrometer, designed to measure lunar surface chemical composition, a neutron spectrometer (to measure hydrogen in the regolith and to search for ice in the polar regions), a magnetometer, and radio tracking for gravity field measurements. Because all instruments were non pointing and had 4π fields of view, the resolution of compositional maps is fixed by the orbital altitude. For the gamma-ray and neutron maps, nominal surface resolution is about 100 km. Because the orbital altitude was lowered during the extended mission, some higher resolution data (about 30 km) are also available (Fig. 3).

Of particular interest in the LP data is the distribution of the element thorium (Th) on the Moon. This element tracks the distribution of KREEP (mentioned before). Most of the Th in the upper crust of the Moon is highly concentrated in a large, regional oval centered on Oceanus Procellarum; smaller concentrations are observed in the floor of SPA basin. The Procellarum Th oval is unexplained. It may represent an original heterogeneity in the crust of the Moon (inherited from global differentiation) or, it could be the result of material excavated and thrown across the surface by the impact that created the Imbrium basin. The (lower in magnitude) Th anomaly in the SPA basin floor suggests that the enrichment in Th here may be the result of the exposure of lower crustal material.

Lunar Prospector's neutron spectrometer detected high concentrations of hydrogen at both poles (Fig. 3). In the form of water ice, the latest results from LP show an amount of hydrogen equivalent to about 10 km^3 of ice; the South Pole has slightly more than the North Pole (contradicting early analyses that suggested more at the North Pole). Moreover, the low-altitude (high-resolution) neutron data show that these high concentrations of hydrogen are correlated with the large areas of darkness in the Clementine mosaics (Fig. 3). This result almost certainly means that large quantities of water ice exist in these dark areas, confirming the earlier result of the Clementine bistatic radar experiment.

Measurements of the magnetic field of the Moon from LP has confirmed that the Moon possesses a small, metallic core, about 400 km in diameter, or, about 2% of the mass of the Moon. This core is mostly iron but may contain significant amounts of iron sulfide (FeS). Numerous, intense zones of magnetism are associated with bright swirl material on the Moon. LP found that some of these magnetic anomalies are intense enough actually to deflect the solar wind from the surface. If such fields are geologically old, there should be enhanced solar wind gas implantation along the margins of these anomalies and shielded areas beneath the magnetic "bubbles."

At the end of its mapping mission, the LP spacecraft was deliberately crashed into the Moon, near the South Pole, in hopes that a cloud of water vapor

might be released, which could then be seen by telescopes on Earth. This experiment was conducted on 31 July 1999; no vapor cloud was detected. This negative result does not mean there is no water ice on the Moon; it only means that we did not detect it in this experiment.

Future Human Activities on the Moon

The discovery of ice on the Moon has enormous implications for a permanent human return to the Moon. Water ice is made up of hydrogen and oxygen, two elements vital to human life and space operations. Lunar ice could be mined and disassociated into hydrogen and oxygen by electric power provided by solar panels deployed in nearby illuminated areas or by a nuclear generator. This hydrogen and oxygen is a prime rocket fuel, giving us the ability to refuel rockets at a lunar “filling station” and making transport to and from the Moon more economical by at least a factor of 10. Additionally, the water from lunar polar ice and oxygen generated from the ice could support a permanent facility or outpost on the Moon. The discovery of this material, rare on the Moon but so vital to human life and operations in space, will make our expansion into the solar system easier and reaffirms the immense value of our own Moon as the stepping-stone into the Universe.

The Moon as a Planetary Touchstone. Beyond obtaining new samples, emplacing a global network of geophysical stations would help us learn more about the Moon’s mantle and core structure, variations in its crustal thickness, and the enigmatic lunar paleomagnetism. It would also lead to a more accurate determination of the enrichment of the Moon in refractory elements by measuring lunar heat flow (because two of these elements, uranium and thorium, are radioactive and thus important sources of heat).

Eventually, humans will probably go to the Moon to live, and establishing a permanent presence there opens up scientific vistas that are difficult to foresee clearly. Each Apollo mission provided some geologic surprise within its sample collection. So there is little doubt that both the variety of rock types and geologic processes that have operated on the Moon exceed by far those that we have currently deciphered. From a permanent lunar base, we could begin a detailed exploration of our complex and fascinating satellite that could last for centuries—and uncover its secrets and also the early history of our home planet as well.

As fascinating as it is, our views on the evolution and history of the Moon have more relevance than just to lunar study. During the last 20 years, we made our first exploration of the planets. We surveyed and photographed all of the terrestrial planets, Mercury, Venus, and Mars, and conducted our initial reconnaissance of the rocky and icy satellites of the giant outer planets. We landed robot spacecraft on Mars and analyzed its surface materials. All of the planetary bodies studied to date show, to different degrees, the same kinds of surface and geologic processes first recognized and described on the Moon. Much of our understanding of planetary processes and history comes by comparing surface features and environments among the planets. In any such comparison, reference is inevitably made to knowledge we obtained from lunar exploration.

One of the most startling results from Apollo was the concept of the magma ocean. The Moon is a relatively small object, transitional in size between planets and asteroids. In general, the amount of heat that a planetary object contains is related to its size; larger planets contain more heat-producing elements. If a body as small as the Moon could undergo global melting, it is a near certainty that the other terrestrial planets also melted. The idea that early Earth underwent global melting has been bandied about for many years; the evidence of the magma ocean made such speculation respectable. Now, we think that early planetary melting may be a widespread phenomenon and could be responsible for creating all of the original crusts of the planets.

Knowing that global melting occurred is one thing, understanding how it operated in detail is another task altogether. The Moon is a natural laboratory to study this process. One of the most fundamental discoveries of Clementine is that the aluminum-rich, anorthosite crust is indeed global and provides strong support for the magma ocean. Our next task is to understand the complex processes at work in such an ocean. Did a “chilled” crust form and if so, are any pieces of it left? Such material would allow us to determine *directly* the bulk composition of the Moon, a parameter that is estimated indirectly (and very imprecisely) now. Are there any highly magnesian rocks in the highlands that are related to the magma ocean, not to the younger magnesium-rich suite of rocks? We have searched the Apollo collections for such rocks but have found none. They may exist at unvisited sites on the Moon.

Another process common to all of the planets is volcanism. The Moon is the premier locality to study planetary volcanism. The flood lavas of the maria span more than a billion years of planetary history and probably come from many different depths within the Moon. Thus, the lava flows are actually probes of the interior of the Moon, both laterally (across its face) and vertically (throughout the depths of the mantle). The inventory and study of the mare basalts will allow us to categorize both of these dimensions and the important additional dimension of *time*. By sampling, chemically analyzing, and dating many different samples of lava that cover the globe, we can piece together the changing conditions of the deep mantle across long periods.

The styles of eruption responsible for the maria appear to be typical of those on other planets. Flood volcanism—the very high rates of effusion responsible for the mare lavas—is seen on every terrestrial planet and appears to be especially widespread on Mars and Venus. What is largely unknown is the size and shape of the vents through which these lavas were extruded. On the Moon, eruptive vents might be exposed in several locations, including within the walls of grabens and irregular source craters. Detailed exploration and study of such features would help us understand a style of volcanism ubiquitous on the planets. Small, dome-like volcanoes, such as found in the Marius Hills, can be explored and examined to understand the styles and rates of lava eruption in creating these features. Small shield volcanoes are common on the surfaces of Mars and, particularly, on Venus. Relatively exotic processes, such as the erosion of terrain by flowing lava, has been proposed for the sinuous rilles of the Moon. Study of large rilles could help us decide whether this concept is correct.

From studying the Moon, we know that impact is one of the most fundamental of all geologic processes. Based on its population of craters of all sizes,

where better to study and understand this important shaper of surfaces than on the Moon? Our ignorance is particularly vast for craters at the larger end of the size spectrum. Craters such as Copernicus (100 km in diameter) offer a window into the upper crust through the study of their ejecta, and a view of the middle level of the crust through their central peaks (which have uplifted rocks from 15–20 km deep). The ubiquity of craters that have such central peaks allows us to reconstruct the nature of the crust in detail. Studying large craters will also clarify the nature of the process of impact. We suspect that large craters grow *proportionally*, that is, they excavate amounts of material to depths that can be predicted from studying smaller craters. But we are not certain that this is true. By studying craters on the Moon, we can determine whether this pattern of growth behaves as predicted.

The giant basins of the Moon pose many mysteries. Understanding such craters is important because we have found basins on the other planets, particularly on Mars and Mercury. Basins form in the earliest stages of planetary history. They excavate and redistribute the crust, serve as depressions where other geologic units may be deposited (such as stacks of thick lava), and may trigger the eruption of massive floods of lava. Yet for all of their importance, we still do not fully understand how far or how deep excavation extends, how the multiple, concentric rings are formed, how the ejected material behaves, and how far ejecta gets thrown as a function of mass. The Moon has preserved more than 40 of these important features for our study, all in various states of preservation and all dating from the very earliest phase of planetary history. Here we can study the process of large-body impact better than anywhere else in the solar system.

The temporal record of impacts in the Earth–Moon system can also be read on the lunar surface. On the basis of evidence of mass extinctions on Earth and from the ages of impact melts, cycles of bombardment and an early impact “cataclysm” have been proposed. Neither of these ideas have been proven, but both are potentially revolutionary and make us look at the history of the planets in a new way. The evidence to test these two ideas lies on the Moon. Episodic bombardment can be tested by sampling the melt sheets of many different craters and dating the samples. Episodes of intense cratering will be evident if groups of melt rocks have the same ages, spaced at constant intervals. On the other hand, a continuous distribution of ages would argue that such episodes of intense cratering do not occur. We cannot conduct this experiment on Earth or on any other planet—this highlights the uniqueness of the Moon for answering many questions in planetary science, questions that have application to a host of other scientific fields. The cataclysm is important because if the Moon underwent such an unusual bombardment history, Earth may also have experienced it, and applying the inferences made from lunar cratering to the other terrestrial planets would have to be reevaluated.

Studying the regolith will be one of the most important tasks during a lunar return. The regolith contains exotic samples flung from rock units hundreds of kilometers away. Using the regolith as a sampling tool, we can conduct a comprehensive inventory of the regional rock units by collecting many samples from the regolith at a single site. The regolith also contains a record of the output of the Sun during the last 3 billion years. To read this record, we must understand how the regolith grows and evolves. This knowledge will come only when we can

study the regolith and its underlying bedrock in detail; to learn how its layers are formed; how the soil is exposed, buried, and re-exposed; and how volatile components might be mobilized and migrate through the soil. This knowledge is essential if we are to realize the goal of using the regolith as a recorder of the solar and galactic particles that have struck the Moon during its history.

Astronomy from the Moon. One day, the Moon will become humanity's premier astronomical observing facility. Consider its advantages. The Moon rotates very slowly (once every 29 days), so its "nighttime" is 2 weeks long. Moreover, because the Moon has no atmosphere, we can observe stars constantly, even in the daytime! The lack of an atmosphere also means that telescopes on the Moon will not be plagued by the "blurring" that a turbulent, thermally unstable air layer causes and that observations will not be degraded by "light pollution," the airglow that interferes with astronomy on Earth. The vacuum of the Moon also means that there are no "absorptions" to prevent certain wavelengths of radiation from being observed, such as the infamous "water absorption" in the atmosphere of Earth. The surface materials could be used as construction material for observatories.

Telescopes in Earth orbit or elsewhere in deep space also have many of these advantages. Why is the Moon better than these localities? The principal reason is that the Moon provides a quiet, stable platform. Seismic activity on the Moon is roughly one million times less than that of Earth. Because the Moon is a primitive, geologically dead world, it does not have the shifting, massive plates of our own dynamic Earth, along with its associated seismic trembling. Such stability of the surface would allow us to construct extremely sensitive instruments for observation, instruments that could not be constructed on Earth.

Likewise, a space-based telescope, such as the Hubble Space Telescope, must have its attitude carefully stabilized to achieve high-resolution observations. In addition, space telescopes have stringent pointing requirements and must not be pointed anywhere near the Sun. Both of these drawbacks mean that a telescope free-flying in space must carry attitude control fuel, precision gyroscopes, and equipment to protect the telescope's optics from solar burn damage. On the Moon, the quiet, stable base of the surface would alleviate such problems, allowing sensitive instruments to be erected and operated easily (7).

One such instrument, an interferometer, consists of an array of smaller telescopes. The smallest object that a telescope can see clearly is directly related to the size of its *aperture*, or the diameter of its mirror or lens. A telescope that has a larger aperture can resolve smaller or more distant features than a telescope that has a smaller aperture. However, there is a practical limit to the size that we can make telescopes. After a certain size is attained, such instruments become unwieldy and unstable. Interferometry is a technique whereby a series of small telescopes are operated as a larger aperture instrument. Each element of the array images some distant object. The light waves from this image are "added" in perfect phase and frequency to identical images obtained from other telescopes in the array, each separated by as much as several kilometers. The effect of this addition is to create an image of the same quality as if a telescope that had an aperture size equal to the separation distance had been used to image the star. This means that we can construct "telescopes" whose effective aperture sizes are *kilometers* across!

Even a small interferometer on the Moon would exceed the resolving capabilities of the very best telescopes on Earth and could even surpass the capabilities of the Hubble Space Telescope. Using such an instrument, we could resolve the disks of distant stars and observe and catalog “star spots,” which are clues to the internal workings of stars. We could see individual stars in distant galaxies and catalog the stellar makeup of a variety of galaxy types. Optical interferometers could look into a variety of nebulae and observe the details of new stars and stellar systems in the very act of formation. Such a window onto the Universe is very likely to revolutionize astronomy in the same way it was changed when Galileo turned his crude “spyglass” toward the heavens in 1609.

The field of planetary astronomy would be completely changed by lunar observatories. The incredible resolving power of these instruments would allow us to examine deep sky systems, resolve the disks of extrasolar planets, and catalog the variety possible in other solar systems of our galaxy. All of our concepts of the way planets are created and evolve are derived from a single example, our own solar system. By observing the vast array of planetary systems circling nearby stars, we could see how they differ in such aspects as the number and spacing of planets, the ratio of giant gas planets to rocky “terrestrial” objects, and the evolution of those individual planets. Spectroscopic observations of these planets would allow us to determine the composition of their atmospheres, if any, and the surface composition, if visible. The compositions of planetary atmospheres could indicate the presence of life on these bodies. Carbon dioxide mixed with free oxygen in a planetary atmosphere would be a tell-tale indication of plant life, which “breathes” the former and manufactures the latter.

Astronomers look at the sky in many wavelengths other than the optical band. High-energy regions of the spectrum, such as X-ray and gamma-ray radiation, also contain important information about processes that occur in stars and galaxies. Supernovas (the sudden explosion of certain stars) produce copious amounts of high-energy radiation and energetic particles, such as cosmic rays. We think that certain particles derived from supernova explosions predate our solar system and that these stellar eruptions can induce planetary formation. We have already mentioned the use of the regolith as a recorder of energetic particle events. Using a high energy lunar observatory, we can watch the effusion of these particles and radiation as they happen (supernovas are common and typically are occurring in some part of the sky at any given time). We have only begun to observe such stellar explosions from space and no doubt have much to learn.

At the other end of the spectrum, the Moon is an ideal place for observation in the thermal infrared and radio bands. Observing the sky in the long wavelengths of the thermal infrared (10 microns and longer wavelengths) is difficult because such detectors measure heat and they must be cooled to very low temperatures for use. Usually, this is done at great cost and difficulty by using cryogenic gases, such as very cold liquid helium (-300°C). Preserving such cold temperatures requires a lot of electrical power. The Moon is naturally cold. Surface temperatures during the lunar night may become as low as -160°C . In the shadowed areas near the South Pole (Fig. 3), it may be as cold as -230°C , only

40 K above absolute zero! These temperatures would permit passive cooling of infrared detectors, allowing telescopes to be operated without costly and difficult-to-use cryogenic, cooling gas. Observing the thermal infrared sky would tell us much about dust clouds and nebulae in which new stars and planets are being formed.

The sky at certain radio wavelengths is almost completely unknown. Earth has an ionosphere, a layer of electrically charged atoms that makes certain radio waves bounce off it, and we cannot see the radio sky at certain frequencies. Moreover, the electrical din of Earth caused by radio stations, microwave cookers, automotive ignition systems, and the thousands of other static generators of modern civilization cause radio astronomers great vexation in their attempts to map the sky. The far side of the Moon is the *only* known place in the Universe that is permanently shielded from the radio noise of Earth. Locating a radio telescope on the far side would permanently place 3600 kilometers of solid rock between the observatory and the radio din of Earth. We will see sky for the first time at some radio wavelengths. History has shown us that whenever we look at the Universe with a new tool or through a new window of frequencies, we learn new things and re-examine present knowledge with new and sometimes startlingly different appreciation.

We can use the distinctive lunar terrain to our advantage. The small, bowl-shaped craters of the Moon are natural features that could be turned into gigantic “dish” radio antennas by laying conductive material (e.g., chicken wire) on their floors and hanging a receiver over and above the center of the crater at the “focus” of the dish. This technique has already been done on Earth at the famous Arecibo Radio Observatory in Puerto Rico, using a natural depression in the limestone bedrock to create a giant “dish” antenna. Interferometers could also be built at radio wavelengths, creating radio telescopes that have huge apertures. The large, flat mare plains would make an ideal site to lay out arrays of smaller telescopes. Manufacture of antenna elements from local resources could make constructing of extremely large instruments feasible.

Using the Moon as an astronomical observatory has great advantages, and many astronomers have taken up the banner for a return to the Moon. In the minds of some, astronomy is the principal reason for a lunar return. However, an observatory on the Moon also has its problems. The ubiquitous and highly abrasive dust must be very carefully controlled. Movements of people and machines will have to be minimized around telescope facilities because the slightest stirring up of dust could coat delicate optical surfaces. We would have to shield energetic detectors carefully from solar flares (this could be done by using the local regolith material.) We must guard against radio contamination of the far side because extensive operations of a base could ruin certain radio astronomical observations. Our task before a lunar return is to understand the impact of each problem fully and devise methods of working around it.

Despite these problems, the Moon offers unique opportunities for astronomy. Each time we see the sky more clearly or more completely, we obtain new insights into the way the universe works. A lunar window on the Universe around us will give us a new appreciation and understanding of both the Universe and of our place in it.

BIBLIOGRAPHY

1. Hartmann, W.K., R.J. Phillips, and G.J. Taylor (eds). *Origin of the Moon*. Lunar and Planetary Institute Press, Houston, TX, 1986.
2. Wilhelms, D.E. *To A Rocky Moon: A Geologist's History of Lunar Exploration*. University of Arizona Press, 1993.
3. Warren, P.H. The magma-ocean concept and lunar evolution. *Ann. Rev. Earth Planetary Sci.* 13: 201–240 (1985).
4. Spudis, P.D. *The Once and Future Moon*. Smithsonian Institution University Press, Washington, DC, 1996.
5. Nozette, S., P. Rustan, L.P. Pleasance, D.M. Horan, P. Regeon, E.M. Shoemaker, P.D. Spudis, C.H. Acton, D.N. Baker, J.E. Blamont, B.J. Buratti, M.P. Corson, M.E. Davies, T.C. Duxbury, E.M. Eliason, B.M. Jakosky, J.F. Kordas, I.T. Lewis, C.L. Lichtenberg, P.G. Lucey, E. Malaret, M.A. Massie, N.H. Resnick, C.J. Rollins, H.S. Park, A.S. McEwen, R.E. Priest, C.M. Pieters, R.A. Risse, M.S. Robinson, R.A. Simpson, D.E. Smith, T.C. Sorenson, R.W. Vorder Brugge, and M.T. Zuber. The Clementine Mission to the Moon: Scientific overview. *Science* 266: 1835–1839 (1994).
6. Spudis, P.D., R.A. Risse, and J.J. Gillis. Ancient multi-ring basins on the Moon revealed by clementine laser altimetry. *Science* 266: 1848–1851 (1994).
7. Burns, J.O., N. Duric, G.J. Taylor, and S.W. Johnson. Observatories on the Moon. *Sci. Am.* 262 (3): 42–49 (1990).
8. Wood, J. Moon over Mauna Loa: A review of hypotheses of formation of Earth's Moon. In W.K. Hartmann, R.J. Phillips, and G.J. Taylor (eds). *Origin of the Moon*. Lunar and Planetary Institute Press, Houston, 1986, pp. 17–55.

PAUL D. SPUDIS
Lunar and Planetary Institute
Houston, Texas

**MUSCLE LOSS IN SPACE:
PHYSIOLOGICAL
CONSEQUENCES****Introduction**

Recent findings clearly suggest that one of the hallmarks of exposing humans and animals to the environment of spaceflight involves the process of muscle atrophy and associated deficits in one's performance capability. This article describes the effects of spaceflight and/or states of muscle unloading on (1) the intrinsic structural and functional properties of skeletal muscle, (2) the performance of muscular activities associated with locomotion and other motor tasks both during and after exposure to spaceflight, and (3) theoretical concepts for enhancing our current understanding of the effectiveness of countermeasures routinely used either to maintain or prevent observed muscle loss and corresponding deficits in musculoskeletal function. Data from both human and animal models from in-flight and ground-based experiments are examined.

Concept of Muscle Homeostasis. The morphological and functional properties of skeletal muscle of adult mammalian species, including humans, normally are quite stable in the gravitational environment of Earth's surface (1 g). Although all of the proteins that make up a muscle undergo continuous synthesis and degradation, the kinetic properties of these pathways are such that muscle mass and protein phenotype, the key properties of strength, endurance, and locomotor/movement capacity, are readily maintained under normal circumstances. However, in the absence of a 1-g stimulus, these homeostatic properties are altered so that the ratio of protein synthesis to degradation is reduced and the ability to maintain protein pools and phenotypes is compromised, thereby contributing to reduced capacity of the muscle system to function at 1 g. Thus, the central objective of any countermeasure strategy during spaceflight is to maintain or closely preserve the neuromuscular properties that exist at 1 g.

Concept of Motor Control of Muscular Activity. Under normal environmental conditions, the ability of the central nervous system (CNS) to control precisely a wide assortment of movement patterns is remarkable, given the variety of conditions under which this control must be managed. All physiological systems that play important roles in the control of movement are affected in one or more ways by the space environment; the chief of them involves a state of unloading on the organism. The supraspinal and spinal pathways of the CNS must accurately and rapidly coordinate a number of neurosensory and neuromotor activities under constantly changing environmental conditions, that is, changing the orientation of the body relative to a 1-g vector on Earth or in a 1-g versus, 0-g environment. In considering any countermeasure to minimize deadaptation to 1 g during spaceflight, it is important to recognize that spaceflight results in changes in the visual, vestibular, and proprioceptive functions; each of them probably contributes to changes in the coordination of activity of the flexor and extensor musculature of the arms and legs in controlling body orientation and locomotion. For example, in microgravity, the relative loadings of both the upper and lower extremities are markedly reduced, thereby affecting the velocity and extent of movement in performing flexion and extension at various joints. Further, the control of movement is shifted toward the upper extremities, hands, and fingers while relying on the lower extremities chiefly for adjusting the center of gravity (1,2). Changes will occur in the neural control of the percentage of the various motoneurons (defined as a nerve fiber arising from the spinal cord that innervates many fibers comprising a given muscle) that are recruited and the force generated by a given muscle or muscle group. This adjustment is accomplished by modulating the number of motoneurons needed for a specific task in accordance with the size principle (3), that is, smaller, high oxidative motor units recruited first and larger low oxidative motor units recruited last, and by frequency modulation.

Associated with the changes in motor activity during spaceflight, as noted above, is the occurrence of a range of detrimental effects on the functional and morphological properties of the skeletal musculature. Changes in the metabolic and mechanical properties of the musculature can be attributed largely to the loss of and alteration in the relative proportion of specific protein systems in the muscles, particularly in those muscles that have an antigravity function at 1 g. These adaptations can degrade performance of routine or specialized motor

tasks, both of which may be critical for survival in an altered gravitational field, that is, during spaceflight and during the return to 1 g. For example, a loss in extensor muscle mass will require a higher percentage of recruitment of the motoneurons for any specific motor task involving extension. Thus, a faster rate of fatigue will occur in activated muscles because, for the same motor task, more motor units will be recruited, thus involving larger, less oxidative motor units that are more fatigable. For this fact alone, it would be advantageous to minimize muscle loss during spaceflight, at least in preparation for the return to 1 g after spaceflight. This section of the article illustrates the inevitable interactive effects of neural, muscular, and endocrine systems in adapting to spaceflight. Only modest progress has been made in understanding the physiological and biochemical stimuli that induce neuromuscular adaptations in a fully integrated system that considers both biological and environmental factors.

Effects of Microgravity and Simulated Models on Skeletal Muscle Properties

Observations on Animals. Numerous effects of altered loading states on the morphological, functional, and molecular properties of mammalian skeletal muscle have been observed. Both ground-based models (e.g., hindlimb unloading) and spaceflight missions (Cosmos Missions and Space Life Sciences 1 and 2) have been used to alter the amount of weight bearing chronically imposed on the muscle(s), thereby changing the amount and type of protein that is expressed in the targeted muscles. These data have been reviewed in detail recently (1–7). A number of adaptations occur as a result of changing the load-bearing function. Briefly, these include the following:

1. Atrophy of both slow-twitch and fast-twitch fibers comprising ankle and knee extensor muscles used for both weight bearing and locomotion (8,9).
2. A change in the type of contractile protein that is expressed in a select population of fibers reflecting a faster phenotype for controlling both cross-bridge and calcium cycling processes, the primary pathways for energy consumption in performing mechanical activity (10–14).
3. Corresponding changes in the functional properties of the muscle manifesting a faster rate of shortening and briefer periods for twitch contraction and relaxation (10,15,16).
4. A reduction in both the absolute and relative force and power generating properties of antigravity and locomotor muscles (10,16).
5. A shift in the force-frequency patterns of antigravity muscles, whereby a higher frequency of electrical stimulation (i.e., action potential frequency) is needed to generate a given submaximal level of force output (10,16–18).
6. A shift in the intrinsic ability of the muscle to use different substrates for energy, whereby the ability to oxidize long chain fatty acids is reduced relative to that of carbohydrates (19).
7. An increase in enzymatic activities supporting the pathways of glycogenolysis and glycolysis (9,12,13,20).

8. Corresponding decreases in the ability of muscle groups to sustain work (i.e., increased fatigability) most likely due to a reduction in the balance of energy supply to energy demand within a given motor unit and a demand for expanded recruitment of faster motor units with less resistance to fatigue (16).
9. A reduction in the oxidative properties of large dorsal root ganglia cells and small motoneurons that could impact the function of sensory neurons and motoneurons (21–23).

Observations on Humans. Based on more than 40 years of experiments in the Russian space program, as well as in the U.S. space program, including the Apollo and numerous Space Shuttle flights, a number of observations have been made that demonstrate a wide range of effects on the motor performance of humans (see Refs. 1 and 2 for recent reviews). Some of these effects include the following:

1. Loss of skeletal muscle mass, particularly in those muscles groups that function at 1 g to maintain extension against normal gravitational loads, often referred to as antigravity muscles (24,25).
2. Reduced ability to exert maximum torque at varying velocities of movement, but particularly at the lower velocities of movement (26,27).
3. Reduction in the size of slow and fast muscle fibers (28).
4. Increase in the proportion of fast myosin in some fibers within 2 weeks of exposure to spaceflight (28,29).
5. Diminished ability to maintain a stable standing posture (26,27).
6. An imbalance of the relative bias of activation of flexor and extensor muscle groups, with greater bias toward flexion in 0 g (30–32).
7. Reduced threshold of the stretch reflex combined with reduced sensitivity to stretch, that is, less gain in responsiveness at a given level of stretch (27).
8. Modified sensitivity to cutaneous vibrations to the sole of the foot (27).
9. Increased susceptibility to fatigue during a given motor task upon return from 0 g to 1 g (1,2).
10. Altered perception of postural position at 0 g and upon return to 1 g (30–32).

Functional Significance of Neuromuscular Adaptations to Spaceflight

Movement Patterns in a Microgravity Environment. When humans enter a microgravity environment, there is an immediate and dramatic reduction in the activation of the extensor musculature required to maintain an upright posture at 1 g (30). The electrical activity (electromyography, EMG) of flexor and extensor muscles in the resting position of the neck, trunk, hip, knee, and ankle reflects a generalized flexor bias during flight compared to 1 g. This bias has been observed during spaceflight when astronauts have been asked to stand upright. This flexor bias effect is independent of whether or not their feet are anchored to

a surface (30). Further, when the astronauts are asked to stand erect with a few degrees of forward tilt, the magnitude of the forward tilt may be as much as four times greater ($\sim 12^\circ$ versus 3°) at 0 g than at 1 g, indicating a relative decrease in extensor activity and/or increase in relative flexor activity. The sites and kinds of sensory information that trigger this exaggerated forward tilt are not understood. Even after returning to 1 g, this residual flexor bias provides a clear indicator of a general adaptation strategy for organizing movements in a 0-g environment.

Although a flexor bias persists during flights, even after adaptation to 0 g, the activity levels of some of the extensor muscles progressively increase within a few days of continued exposure to the 0-g environment (30). This recovery of extensor activity and continued elevation of flexor activity has been clearly documented in ground-based models of weightlessness. For example, extensor EMG activity essentially disappears immediately upon unloading of the hindlimbs in rats (33). Within hours, however, some EMG activity reappears during continued hindlimb unloading, and by 7 days, the total daily amount of activation is near normal levels. This pattern has been observed in both predominantly slow (e.g., the soleus) and fast (e.g., the medial gastrocnemius) ankle extensors. In contrast, the EMG activity of the tibialis anterior, an ankle flexor, is significantly elevated throughout the unloading period (33). The "recovery" to normal or near normal levels of extensor EMG activity while remaining "unloaded" suggests that the CNS is "programmed" so that general extensor bias continues as it does at 1 g under normal gravitational loads (33). This apparent residual bias may have been permanently acquired during development as a result of the daily sensory cues of a 1-g environment. Alternatively, this extensor bias could be inherent in the design of the CNS, that is, independent of any activity-dependent events associated with movement control in a 1-g environment.

Movement Characteristics at 1 g After Exposure to 0 g. The ability to perform movements, including posture and locomotion at 1 g, is adversely affected by exposure to as little as 1 week of spaceflight (26). All crew members tested to date have experienced some postural instability for 1–2 weeks following spaceflight, or even longer in some instance. This instability, which varies markedly from crew member to crew member, reflects alterations in perception, sensitivity, and responsiveness. In many cases, the altered motor functions may not be readily apparent due to use of compensatory mechanisms such as maintaining a wider stance, taking shorter steps, having a greater dependence on visual cues, and generally being more cautious.

Other physiological measures also reflect altered movement control following spaceflight, particularly that of altered postural responses to horizontal perturbations, for example, unusual magnitudes and durations of activation of extensor and flexor motor pools. Based on studies of the visiting crews in the Salyut-6 missions, ranging in duration from 4–14 days (most for 7 days), the EMG response of the soleus and tibialis anterior muscles to perturbations of the standing position was almost doubled, and the response time to the perturbation was three times longer after than before flight (27). Severe postural disruptions following 4–10 days of spaceflight on the Shuttle have also been reported (27). A rapid recovery rate was evident immediately after flight, and most of the recovery occurred within the first 10–12 hours postflight, followed by a slower recovery

during the next 2–4 days. Further, it was estimated that 50% of the recovery occurred within 3 hours postflight. Adverse postural effects, however, persisted for as long as 42 days after a 175-day flight.

As was true for performance in a maximum torque–velocity test of the plantarflexors (e.g., the muscles that lift the heel off the ground), the duration of the spaceflight has not proven an important determinant in the severity of postural stability (2,34). For example, cosmonauts that had been on the Mir station for 326 days had a similar EMG amplitude response to postural perturbations immediately after flight as before flight. In contrast, the EMG response was doubled in cosmonauts from either a 160-day or a 175-day spaceflight (34). However, the similarities in the magnitude of the neural control adaptations may reflect improvements in the methods for countering the degradation in movement control, as the flights have become longer.

Another clear example of the modification of the input–output ratio of the motor system was demonstrated after 7 days of dry immersion. Before immersion, subjects were able to increase the force in relatively constant increments up to about 50% of maximal voluntary contraction in a succession of about 10 trials. After flight, the subjects overestimated the target force considerably even at the lower force levels, and the force differential became even more distorted at higher torques (27).

Although many adaptations are clearly manifested during the performance of motor tasks at 1 g after having adapted to spaceflight conditions for a specific duration, many details regarding the specific adaptations to spaceflight are not available because of the difficulty in conducting well-controlled experiments in very complex space mission flight plans and objectives. All studies of humans reflect some unknown combination of the effects of spaceflight conditions, the measures used to counter spaceflight effects, and individual differences in responsiveness to both spaceflight and the countermeasures used.

Changing the magnitude of the gravitational vector also alters body movement perception. There are immediate and longer term effects of these altered forces on perception. It is clear, for example, that there are disturbances in oculomotor control, vestibular function, pain sensitivity, muscle stretch sensitivity, joint position sense, and cutaneous sensitivity to vibration, all of which may play some role in modifying motion–position perception in response to spaceflight (2,27,35). Altered perceptions of speed of movement, the effort it takes to perform a movement, and movement of the body relative to its surroundings have been reported when alternating periods of 0 g and 2 g are imposed during parabolic flights (36). When subjects raised or lowered their bodies, from a squat position during the 2-g phase of parabolic flight, perceptual distortions of movement were evident. These findings were interpreted to indicate that the motor control of skeletal muscles had been calibrated to a 1-g reference level and that these illusions resulted from mismatches between the efferent control signals and the expected patterns of associated spindle activity.

Perception of upper limb position was also studied during parabolic flights, with and without vibration of the biceps or triceps brachii tendons. Because tendon vibration activates sensory Ia neurons of muscle spindles and to some extent the sensory type IIb neurons (e.g., afferent nerve fibers) from the muscle tendon organs and because these receptors it is thought, contribute to the

sensation of angular displacement, these studies provided some insight into their potential role of proprioception (i.e., how the body senses its position on the ground and in space) in the diminished accuracy of position sense observed after spaceflight (36–38). The perceived magnitude of displacement from the apparent limb position upon tendon vibration was $1.8g > 1g > 0g$. The subject's perceptions of displacement were consistent with the actual displacements. The results of these experiments suggest that higher g forces on the body required greater postural tonus and an altered gain for the muscle spindles (38).

Many astronauts have reported that they are not aware of the positions of their limbs when they shut their eyes to sleep or relax while weightless; one crew member stated, "It is almost as if the limbs are gone" (39). When they tense their muscles, the position sense returns. One explanation for this phenomenon is not that sensitivity of the receptors is blunted, but that the stimulus is reduced. Interestingly, this sensation is the antithesis of the phantom-limb phenomenon described by amputees, that is, sensing the presence and even the movement of a limb, even though the limb has been amputated. A consistent observation by those who have experienced 0g for prolonged periods has been a sensation of heaviness of the body, particularly of the head, upon return to 1g. The selective atrophy of the extensor musculature associated with weightlessness could contribute to this sensation of heaviness. Muscle atrophy is not the only factor, given that this sensation of heaviness disappears in some crew members within a few hours. Performing a task with atrophied muscles will require a sense of greater effort than normal because more motor units must be recruited at a higher frequency of activation. In turn, this elevated recruitment is likely to increase the activation of muscle proprioceptors, which are the sensor organs that detect mechanical stress on the muscle.

Contractile Properties of Skeletal Muscle Following Spaceflight. The ability to produce maximum force-producing ankle rotation, which is also called plantar flexor torque, at velocities ranging from 0 to $180^\circ/s$ is reduced in short-term (7–14 days) and long-term (75–237 days) spaceflight. After short-term spaceflight, maximum isometric torque decreased by 18% and at $60^\circ/s$ by 38%. After long-term spaceflight, maximal torques were 25, 12, 10, and 18% lower than preflight values at 0, 60, 120 and $180^\circ/s$, respectively. The changes in torque among the cosmonauts varied considerably, ranging from -60 to $+15\%$ of preflight values. Although muscle atrophy almost certainly contributes to the reduced torques commonly observed after short- and long-term spaceflights, neural activation of the motoneurons innervating the muscles must also be affected. The highly variable losses in torque, at high speeds in some cases and at low speeds in others, cannot be easily explained by changes in muscle properties alone (26,27).

Some of the adaptations in the motor responses noted above may reflect, at least in part, the effects of muscle atrophy. The reduced force potential could exacerbate the postural instability of the astronauts and cosmonauts usually attributed to the neural control system upon return to 1g. This possibility seems particularly feasible because the fibers innervated by the motoneurons that have the larger role in maintaining routine posture (i.e., the slow motor units) are those that often atrophy the most. Further, if the nervous system is not aware of the reduced muscle force potential and does not adjust the output signal accordingly, then the motor output will be reduced. This inappropriate neural input

to output relationship will result in exaggerated movement or sway during standing and may even result in the loss of balance, as described above.

In summary, during and after spaceflight, the effectiveness of the neuro-motor system is clearly compromised. There could be a degradation in the function of the muscles, synapses within the spinal cord, reception of sensory information by the brain and, in some cases, interpretation and perception of the environment. Dysfunction at any one of these levels at some critical time during a flight could have a major impact on the success of a mission and the safety of the crew members.

Causative Factors for Muscle Alterations in Response to Weightlessness

What are (1) the potential stimuli that are likely to induce muscle adaptations to spaceflight and (2) the cellular/subcellular processes involved in the adaptive response? Adaptations at the cellular level that result in a change in the quantity and/or quality or the type of protein expression in response to reduced mechanical activity can be regulated theoretically at several levels of control involving transcriptional, pretranslational, translational, and posttranslational processes (40). In considering the types of adaptations reported above in the rodent model, available evidence suggests that all of these processes are likely to play pivotal role (2,41).

Neuromuscular Activity and Its Role in Plasticity

Hindlimb Unloading. Adaptation to chronic unloading of skeletal muscles is determined in part by the manner in which the muscles are activated. The activity patterns of the rat soleus, medial gastrocnemius, and tibialis anterior muscles have been monitored from the same chronically implanted intramuscular electrodes before and during 1 month of hindlimb unloading; this is a procedure whereby traction of the tail is applied to a rodent so that the hindlimbs can be lifted off of the floor of the cage (33). Total daily EMG activity (mV/s), measured in the soleus and medial gastrocnemius, was significantly reduced on the day of initial unloading, was similar to control levels by 7–10 days of ongoing unloading, and continued at nearly normal levels for the remainder of the experimental period. Daily EMG levels of the tibialis anterior were above normal during all postunloading days. In addition, the interrelationships of the EMG amplitude patterns, a reflection of recruitment patterns, between the soleus and medial gastrocnemius were altered on the day when initiating unloading, but recovered to a normal pattern by day 7 of unloading (42). These data indicate that the neural mechanisms controlling hindlimb muscles, at least for some extensor muscles, initially change in response to the unloaded environment but might return to normal within a short period of time after the unloading began.

However, based on 16 min of continuous EMG activity per day from the soleus on 7 and 4 days before unloading, as well as after 4 and 7 days of unloading, Riley et al. (43) concluded that the average total time of soleus activity (normalized per hour) in suspended rats was about 12%, compared to preunloading values. These authors also concluded that the activity developed a “phasic,” compared to a “tonic” pattern. The amplitude (root mean square) of the

signals was smaller (25–50%) during, compared to before unloading. Blewett and Elder (44) recorded a total of 2.4 h/day (6 s/min) of activity from the soleus and plantaris muscles (both primary ankle extensors) during several days before and after unloading; both overall mean amplitude and frequency of motor unit action potentials (based on a “turns analysis”) were significantly decreased during unloading, compared to normal cage activity. This experiment differed from that of Alford et al. (33) in the total duration of EMG recording and also in the experimental design. For example, rats were allowed to ambulate normally for a 2 h period each week throughout the experimental period, a procedure that may have impacted the results because short periods of load-bearing, it has been shown, have a significant effect on other muscle properties (2,5).

Combined, these data indicate a general decrease in activation duration and amplitude of extensor muscles during the first week of unloading. On the other hand, after longer periods of unloading, extensor EMG activity may return to normal. Although chronic tension levels in muscles of suspended rats have not been recorded, it is reasonable to assume that forces in the plantarflexors were small. In addition, it is likely that the plantarflexors were further unloaded when the limb is in the plantarflexed position, the usual position during unloading. In contrast, loads on the tibialis anterior muscle were most likely to be slightly elevated when the muscle is maintained in a “stretched” position. It seems unlikely, however, that the greater activity (3–4 times) of the tibialis anterior was due to increased stretching of the muscle during unloading, compared to that during routine cage activity, because most muscle stretch receptors accommodate rapidly to sustained stretch. In any case, the effects of unloading on the actual loading properties of muscle need to be substantiated by recording muscle forces pre-, during, and postspaceflight.

Spaceflight. The effects of spaceflight on chronic neuromuscular activity in rats are less substantiated than for hindlimb unloading. Presumably, muscle forces during spaceflight would be minimal, although no force data are available. In humans during spaceflight, the tonic activity (EMG) of the soleus (a plantarflexor) is reduced, whereas the tonic activity of the tibialis anterior (a dorsiflexor) is enhanced during postural adjustments (30). This activity reversal of extensors and flexors normally observed at 1 g has also been reported during parabolic flights in humans (45) and in monkeys during and after short-term flights (46–48). No chronic EMG or muscular force data from either humans or animals during spaceflight have been published.

Reductions in Muscle Strength and Power. Studies of both humans and animals clearly show that the force generating capacity of extensor muscles of the hindlimb are reduced to varying degrees following either spaceflight or hindlimb unloading. This reduction has been attributed to (1) a decrease in muscle mass reflecting a reduction in the cross-sectional area of the muscle fibers, (2) a reduction in the capacity to activate the muscle via supraspinal pathways, and (3) a reduction in the specific force of the muscle (reduced force corrected for fiber cross-sectional area) (2,5,46). Though it is reasonable to conclude that the reduced force due to atrophy is due to a loss in contractile protein that contributes to the contraction process, the underlying mechanisms responsible for the inability to activate the various motoneurons that innervate the muscles are poorly understood. The same holds true for the factors involved in

the reduction in specific tension. Clearly, more extensive research is needed on this important problem.

Mechanisms of Muscle Atrophy. During states of hindlimb unloading, the rate of total protein synthesis (a translational process) is significantly reduced within the first few hours of creating the unloaded state. This is coupled to a subsequent transient increase (during the next several days) in the net rate of protein degradation, thereby resulting in a ~50% smaller protein pool comprising the muscle, that is, the muscle becomes significantly atrophied (7,49). Though both the initiating events and the signal transducing process(es) associated with the atrophy response remain largely unknown (see above), the involvement of either growth factor(s) downregulation or the catabolic actions of other hormones appear to be involved. For example, mRNA signals for insulin-like growth factor-1 (IGF-1) expression in skeletal muscle is reduced in hindlimb muscle when the weight-bearing activity of the animal is reduced (50). In contrast, recent findings suggest an opposite response when a skeletal muscle is functionally overloaded, that is, IGF-1 expression is enhanced (51). Recent studies of cardiac muscle further suggest that IGF-1 expression is linked to the loading state imposed on the system (4). The potential roles of other hormonal factors in modulating the atrophy response are uncertain.

In this context, there appears to be a critical interplay between mechanical factors and growth-stimulation factors, such as growth hormone, in maintaining muscle mass when challenged by a state of unloading (52,53). Furthermore, it has been shown that the time course of the muscle atrophy response to unweighting is altered when cellular glucocorticoid receptors are pharmacologically blocked (54). Under conditions involving muscle wasting in response to exogenous glucocorticoid treatment, a key enzyme, glutamine synthase, is also unregulated by elevations in the circulating level of this hormone (55). This enzyme it is thought, regulates both the formation and release of the amino acid, glutamine, that is, the primary amino acid to which most amino acids are converted during the protein degradation process, from the muscle during the wasting process. Experiments in which the level of glutamine is artificially elevated in both the plasma and muscle during glucocorticoid treatment, markedly reduces both the atrophy process and the decrease in total protein and myosin protein synthesis rates that occur under these conditions (55). On the other hand, agents that, it is thought, inhibit the proteasome axis component in the cascade of protein degradation processes appear to be partially effective in ameliorating the atrophy response to weightlessness (56). Clearly, a more basic understanding of the interactions of hormonal and activity factors in the regulation of protein synthetic and degradation processes, especially in the context of the atrophy response associated with muscle unloading, is needed.

Alterations in Myosin Phenotype. Results from both cardiac and skeletal muscle suggest that transcriptional and pretranslational control of the slow type of myosin heavy chain (MHC) gene, a key gene encoding a regulatory contractile protein, is highly regulated by thyroid hormone (T₃) (40,41). For example, T₃, in conjunction with its nuclear receptor and other nuclear regulatory proteins, acts as a negative modulator of transcription of the beta (slow or type I) MHC gene while concomitantly exerting positive transcriptional control of the cardiac fast, alpha MHC gene (57). Thus, it is interesting that the downregulation of the slow

myosin gene typically seen during states of unloading can be inhibited by making the animals hypothyroid (58). Collectively, these findings suggest that changes in the loading state may alter the muscle's responsiveness or sensitivity to thyroid hormone. Furthermore, recent findings on cardiac muscle also suggest that transcription of the beta (slow, type I) MHC gene can be positively regulated by expression of a nuclear factor(s) that binds to a specific DNA sequence (designated as beta e2) upstream of the gene's transcriptional initiation site (57). This factor can be upregulated in the rodent heart in response to pressure overload. Thus, there appears to be a complex interaction of mechanically (loading state) and hormonally induced transcriptional factors that are involved in regulating MHC plasticity in response to altered states of muscle loading (40). Understanding the regulatory factors associated with slow myosin gene expression is important because it is predominantly the slow myosin isoform that is sensitive to the gravity state. Furthermore, motor units expressing slow myosin are those predominantly recruited for posture control and low intensity movements (59,60).

Perspectives in Preventing Muscle Atrophy and Dysfunction

Basic Physiological Principles for Developing Exercise Countermeasures.

Any exercise-related countermeasure for preserving skeletal muscle function will be manifested via the spinal mechanisms that regulate the order and number of motor units recruited. In essence, all movements represent the net effect of the number of motor units recruited and the combination of motor units for each muscle that will be recruited, combined with the mechanical restraints placed on the muscles. The selection of which and how many motor units will be recruited is defined in some manner when kinematic features of the motor task are selected. Muscle activity also can be imposed by electrical stimulation of its nerve where, unless special technical considerations are instrumented, the most readily stimulated muscle fibers will be those innervated by the largest axons (and thus probably belong to the largest motor units). Such a recruitment order determined largely by axon diameter is opposite to that normally used by the central nervous system. Otherwise, the same general principles described below apply to electrical stimulation of muscle as a potential countermeasure.

In designing an exercise countermeasure, the major variables to modulate are the "level of effort," that is, the number (and frequency to some degree) of motor units recruited and the speed at which the muscle will shorten or lengthen. For any given level of recruitment, the changes in muscle length will be defined by the mechanical conditions under which the motor units are activated. The force produced will be a function primarily of the number of motor units (and thus muscle fibers) recruited and the mechanics that define the velocity and direction of movement. Because the force-velocity relationships are somewhat predictable, the "types" of exercise (high resistance, high power, low resistance, etc.) largely reflect the number of motor units recruited and the temporal pattern for their recruitment. Whether that force is sufficient to shorten or lengthen the muscle and the speed at which the displacement will occur depend on the loading conditions. A high-resistance exercise is one in which a high proportion of the

units within the appropriate motor pools is recruited and a high load is imposed, resulting in a relatively slow velocity of shortening. If the same recruitment pattern occurs and the load is reduced, then the velocity of movement will increase hyperbolically.

Although these basic physiological concepts derived from isolated (e.g., *in situ*) experiments are well recognized and generally accepted, they have not been translated into a rational and systematic approach to developing more effective countermeasures for the neuromotor deficits that develop during prolonged spaceflight. Further, a more integrative rather than reductive approach to motor performance is needed. A more integrative physiological perspective must also be maintained in assessing the metabolic consequences and the corresponding adaptations to spaceflight and exercise. Given the interdependence of the motoneuronal and muscle metabolic properties, however, recruitment and metabolic responses of recruitment are essentially inseparable.

Relevant questions for maintaining normal muscle tissue properties include the following:

1. What combinations of forces and velocities will most efficaciously maintain the normal physiological status for each type of motor unit and muscle fiber type?
2. What are the differences in the responsiveness of fiber types and muscle types to specific muscle force-velocity events? For example, does this responsiveness differ in the arm versus the leg, flexor versus extensor muscles, etc.?
3. What durations and intermittencies, that is, work-rest ratios of the mechanical stimuli, are necessary to maintain the normal properties of a muscle fiber?

An implied assumption in these questions is that there are some mechanical event(s) associated with exercise that produce the necessary stimuli for cell maintenance. However, these stimuli could be metabolic or some other event related to excitation-contraction coupling. In any case, the same considerations of the variables noted above would be appropriate for each potential physiological modulator.

The Quantity of Activity Needed for Muscle Homeostasis. To counter the atrophic effects of spaceflight, one needs to know the means by which the space environment induces flight effects. Two of the prevalent hypotheses are that muscles atrophy during flight because of a reduction in (1) activation of muscles, or (2) the muscle forces associated with the reduction in activation. For example, a common concept that has prevailed for many years is that muscles enlarge when they are active and atrophy when they are inactive. Further, a linear and direct relationship between muscle fiber size and neuromuscular activity or exercise level is often assumed. It is clear, however, that this assumption is incorrect or at best, misleading. For example, within a given muscle, those muscle fibers that are used (i.e., recruited) least often are usually the largest fibers. Analyses of biopsies from endurance-trained swimmers and weight lifters also illustrate that the amount of activity is poorly correlated with fiber size (61).

Thus, it is apparent that the effectiveness of an exercise as a countermeasure to muscle atrophy cannot be based solely on the quantity (total time, number of repetitions, etc.) of exercise. To maintain muscle mass, it appears that a relatively small amount or duration of activity per day is needed and the amount needed varies widely among fiber types and specific muscles. The more important factor appears to be the mechanical load on the muscle during activation (1,5,40). This view certainly appears to be true in hindlimb-suspended rats when the animals are exercised intermittently. These studies suggest that 6 minutes per day of climbing a grid with attached weights (i.e., a relatively high load exercise) had almost the same effect of ameliorating muscle atrophy as 90 minutes of daily treadmill exercise (i.e., a relatively low load exercise) (17). Thus, some minimum amount of muscle activation and force may be required to maintain muscle mass.

In defining exercise protocols and devices to counter the effects of spaceflight on skeletal muscle, the most efficacious exercise may be unique for each muscle group, for example, extensors versus flexors and muscle type (i.e., muscles that are comprised predominantly of slow vs. fast fibers). Further, an exercise regimen that may prevent muscle atrophy may not be the most efficacious in preventing demineralization of bone. However, reasonable compromises in exercise prescriptions during spaceflight can and must be defined so that a crew member will not need to exercise several hours each day to maintain an acceptable functional state while spending prolonged periods in space and during periods of reduced gravitational forces while on the Moon or Mars.

The Impact of Activity–Hormonal Interactions. Neuromuscular activity may play a facilitatory role, rather than a direct role, in maintaining muscle mass. For example, it is increasingly obvious that there can be important interactive effects between exercise and hormones. Glucocorticoids can induce marked and selective atrophy of fast muscles, and weightlifting or treadmill exercise during glucocorticoid administration can greatly reduce the severity of the atrophic response (62). Similarly, growth hormone alone can significantly decrease the severity of atrophy induced by hindlimb unloading of rats (52,53). Interestingly, this effect is greatly amplified when the growth-hormone-treated suspended rats are exercised (climbing a 1-meter grid inclined at 85° with weights attached as little as 15 times/day) (52,53). Additional important aspects of neuromuscular activity and growth factors were discussed before in “Mechanisms of Muscle Atrophy.”

Defining the Acceptable Limits of Muscle Dysfunction in Microgravity.

From an operational point of view, some consensus needs to be formulated regarding how much loss of function can be tolerated without a significant compromise in safety and possible long-term consequences. For example, one 10-min exercise period per day may be sufficient to maintain 90% of normal function of the extensors of the ankle, knee, hip, trunk, and neck, whereas it may require 90 min/day to maintain 95% of normal function. Does 90% of normal function provide an acceptable margin of safety? Similar operational issues are relevant for each physiological system. Individual differences among flight candidates should also be taken into account, in particular, because the results from virtually every study of spaceflight and ground-based models of spaceflight have demonstrated marked individual differences in the response of the

neuromuscular system. These unique individual responses may hold the key to better understanding of the etiology and magnitude of these specific effects. An integrative physiological perspective and experimental approach in determining the adaptability of humans to spaceflight is essential. Considerable insight into this problem of safety factors in biological systems can be gained by applying these concepts outlined by Diamond (63,64).

Concluding Perspectives. As we enter the new millennium, it is most apparent that mankind is set on expanding the human presence further into the universe. These visionary missions will require a tremendous commitment of resources plus the expanded knowledge base that will be necessary to maintain the health and safety of astronauts during missions that may take as long as 4–5 years to complete. This article has focused on only one system, skeletal muscle, but it is apparent that all systems of the body are compromised in the absence of gravity. Therefore, scientists must begin thinking about using an integrative biological approach in putting a strategic plan together to expand our understanding of how man can endure periods of time in a weightless environment that exceed by four- to fivefold the current duration of existence in such a unique environment. The question is, are we up to this challenge?

ACKNOWLEDGMENT

This work was supported by NIH grants 16333 (VRE) and AR 30346 and NSBRI NCC-58-A (KMB).

BIBLIOGRAPHY

1. Edgerton, V.R., and R.R. Roy. Neuromuscular adaptation to actual and simulated weightlessness. In S.L. Bonting (ed.), *Advances in Space Biology and Medicine*, Vol. 4, JAI Press, Greenwich, CT, 1994, pp. 33–67.
2. Edgerton, V.R., and R.R. Roy. Neuromuscular adaptations to actual and simulated spaceflight. In M.J. Fregly and C.M. Blatteis (eds), *Handbook of Physiology. Section 4. Environmental Physiology. III. The Gravitational Environment*, Chap. 32, Oxford University Press, New York, 1996, pp. 721–763.
3. Henneman, E., G. Somjen, and D.O. Carpenter. Functional significance of cell size in spinal motoneurons. *J. Neurophysiol.* 28: 560–580 (1965).
4. Donohue, T.J., L.D. Dworkin, M.N. Lango et al. Induction of myocardial insulin-like growth factor gene expression in left ventricular hypertrophy. *Circulation* 89: 799–809 (1994).
5. Roy, R.R., K.M. Baldwin, and V.R. Edgerton. The plasticity of skeletal muscle: Effects of neuromuscular activity. In J. Holloszy (ed.), *Exercise and Sports Sciences Reviews*, Vol. 19, Williams and Wilkins, Baltimore, MD, 1991, pp. 269–312.
6. Roy, R.R., K.M. Baldwin, and V.R. Edgerton. Response of the neuromuscular unit to spaceflight: What has been learned from the rat model? In J. Holloszy (ed.), *Exercise and Sports Sciences Reviews*, Vol. 24, Williams and Wilkins, Baltimore, MD, 1996, pp. 399–425.
7. Thomason, D.B., and F.W. Booth. Atrophy of the soleus muscle by hindlimb unweighting. *J. Appl. Physiol.* 68: 1–12 (1990).
8. Martin, T.P., V.R. Edgerton, and R.E. Grindeland. Influence of spaceflight on rat skeletal muscle. *J. Appl. Physiol.* 65: 2318–2325 (1988).

9. Miu, B., T.P. Martin, R.R. Roy, V. Oganov, E. Ilyina-Kakueva, J.F. Marini, J.J. Leger, S.C. Bodine-Fowler, and V.R. Edgerton. Metabolic and morphologic properties of single muscle fibers in the rat after spaceflight, Cosmos 1887. *FASEB J.* 4: 64–72 (1990).
10. Caiozzo, V.J., M.J. Baker, R.E. Herrick, M. Tao, and K.M. Baldwin. Effect of spaceflight on skeletal muscle: Mechanical properties and myosin isoform content of a slow muscle. *J. Appl. Physiol.* 76: 1764–1773 (1994).
11. Haddad, F., R.E. Herrick, G.R. Adams, and K.M. Baldwin. Myosin heavy chain expression in rodent skeletal muscle: Effects of exposure to zero gravity. *J. Appl. Physiol.* 75: 2471–2477 (1993).
12. Jiang, B., R.R. Roy, I.V. Polyakov, I.B. Krasnov, and V.R. Edgerton. Ventral horn cell responses to spaceflight and hindlimb suspension. *J. Appl. Physiol.* 73: 107S–111S (1992).
13. Ohira, Y., B. Jiang, R.R. Roy, V. Oganov, E. Ilyina-Kakueva, J.F. Marini, and V.R. Edgerton. Rat soleus muscle fiber responses to 14 days of spaceflight and hindlimb suspension. *J. Appl. Physiol.* 73 (Suppl.): 51S–57S (1992).
14. Talmadge, R.J. Myosin heavy chain isoform following reduced Neuromuscular activity: Potential regulatory mechanisms. *Muscle and Nerve* 23: 661–679 (2000).
15. Fitts, R.H., J.M. Metzger, D.A. Riley, and B.R. Unsworth. Models of disuse: A comparison of hindlimb suspension and immobilization. *J. Appl. Physiol.* 60: 1946–1953 (1986).
16. Winiarski, A.M., R.R. Roy, E.K. Alford, P.C. Chiang, and V.R. Edgerton. Mechanical properties of rat skeletal muscle after hindlimb suspension. *Exp. Neurol.* 96: 650–660 (1987).
17. Herbert, M.E., R.R. Roy, and V.R. Edgerton. Influence of one-week hindlimb suspension and intermittent high load exercise on rat muscles. *Exp. Neurol.* 102: 190–198 (1988).
18. Pierotti, D.J., R.R. Roy, V. Flores, and V.R. Edgerton. Influence of 7 days of hindlimb suspension and intermittent weight support on rat muscle mechanical properties. *Aviation Space Environ. Med.* 61: 205–210 (1990).
19. Baldwin, K.M., R.E. Herrick, and S.A. McCue. Substrate oxidation capacity in rodent skeletal muscle: Effects of exposure to zero gravity. *J. Appl. Physiol.* 75: 2466–2470 (1993).
20. Hauschka, E.O., R.R. Roy, and V.R. Edgerton. Size and metabolic properties of single muscle fibers in rat soleus after hindlimb suspension. *J. Appl. Physiol.* 62: 2338–2347 (1987).
21. Ishihara, A., Y. Ohira, R.R. Roy, S. Nagaoka, C. Sekiguchi, W.E. Hinds, and V.R. Edgerton. Influence of spaceflight on succinate dehydrogenase activity and soma size of rat ventral horn neurons. *Acta Anat.* 157: 303–308 (1996).
22. Ishihara, A., Y. Ohira, R.R. Roy, S. Nagaoka, C. Sekiguchi, W.E. Hinds, and V.R. Edgerton. Effects of 14 days of spaceflight and 9 days of recovery on cell body size and succinate dehydrogenase activity of rat dorsal root ganglion neurons. *Neuroscience* 81: 275–279 (1997).
23. Ishihara, A., Y. Ohira, R.R. Roy, S. Nagaoka, C. Sekiguchi, W. Hinds, and V.R. Edgerton. Comparison of the response of motoneurons innervating perineal and hindlimb muscles to spaceflight and recovery. *Muscle & Nerve* 23: 753–762 (2000).
24. LeBlanc, A., P. Gogia, V. Schneider, J. Krebs, E. Schonfeld, and H. Evans. Calf muscle area and strength changes after five weeks of horizontal bed rest. *Am. J. Sports Med.* 16: 624–629 (1988).
25. LeBlanc, A., R. Rowe, V. Schneider, H. Evans, and T. Hedrick. Regional muscle loss after short duration spaceflight. *Aviation Space Environ. Med.* 66: 1151–1154 (1995).
26. Kozlovskaya, I.B., Yuy. V. Kreidich, V.S. Oganov, and O.P. Koserenko. Pathophysiology of motor functions in prolonged manned space flights. *Acta Astronautica* 8: 1059–1072 (1981).

27. Kozlovskaya, I.B., I.F. Dmitrieva, L. Grigorieva, A. Kirenskaya, and Y. Kreidich. Gravitational mechanisms in the motor system. Studies in real and simulated weightlessness. In V.S. Gurfinkel, M.E. Ioffe, J. Massion, and J.P. Roll (eds), *Stance and Motion: Facts and Concepts*, Plenum Press, New York, 1988, pp. 37–48.
28. Edgerton, V.R., M.-Y. Zhou, Y. Ohira, H. Klitgaard, B. Jiang, G. Bell, B. Harris, B. Saltin, P.D. Gollnick, R.R. Roy, M.K. Day, and M. Greenisen. Human fiber size and enzymatic properties after 5 and 11 days of spaceflight. *J. Appl. Physiol.* 78: 1733–1739 (1995).
29. Zhou, M.Y., H. Klitgaard, B. Saltin, R.R. Roy, V.R. Edgerton, and P.D. Gollnick. Myosin heavy chain isoforms of human muscle after short-term spaceflight. *J. Appl. Physiol.* 78: 1740–1749. (1995).
30. Clement, G., and F. Lestienne. Adaptive modifications of postural attitude in conditions of weightlessness. *Exp. Brain Res.* 72: 381–389 (1988).
31. Clement, G., V.S. Gurfinkel, F. Lestienne, M.I. Lipshits, and K.E. Popov. Adaptation of postural control to weightlessness. *Exp. Brain Res.* 57: 61–72 (1984).
32. Clement, G., V.S. Gurfinkel, and F. Lestienne. Mechanisms of posture maintenance in weightlessness. In I. Black (ed.), *Vestibular and Visual Control On Posture and Locomotor Equilibrium*, Karger, Basel, Switzerland, 1985, pp. 158–163.
33. Alford, E.K., R.R. Roy, J.A. Hodgson, and V.R. Edgerton. Electromyography of rat soleus, medial gastrocnemius and tibialis anterior during hindlimb suspension. *Exp. Neurol.* 96: 635–649 (1987).
34. Kozlovskaya, I.A., V.A. Barmin, V.I. Stepanov, and N.M. Kharitonov. Results of studies of motor functions in long-term space flights. *Physiologist* 33: S1–S3 (1990).
35. Cohen, M.M. Perception and action in altered gravity. In B. Cohen, D. Tomoko, and F. Guedry (eds), *Sensing and Controlling Motion: Vestibular and Sensorimotor Function*, Vol. 656, Annals of the New York Academy of Sciences, New York, 1992, pp. 354–362.
36. Lackner, J.R., and A. Grabiell. Illusions of postural, visual, and aircraft motion elicited by deep knee bends in the increased gravito-inertial force phase of parabolic flight. *Exp. Brain Res.* 44: 312–316 (1981).
37. Lackner, J.R., and M.S. Levine. Changes in apparent body orientation and sensory localization induced by vibration of postural muscles: Vibratory myesthetic illusions. *Aviation Space Environ. Med.* 50: 346–354 (1979).
38. Lackner, J.R., and P. Dizio. Gravito-inertial force level affects the appreciation of limb position during muscle vibration. *Brain Res.* 592: 175–180 (1992).
39. Schmidt, H.H., and D.J. Reid. Anecdotal information on space adaptation syndrome. NASA/Space Biomedical Research Institute and USRA/Division of Space Biomedicine, July 1–21, 1985.
40. Booth, R.W., and K.M. Baldwin. Muscle plasticity: Energy demand and supply processes. In L.B. Rowell and J.T. Shephard (eds), *American Physiological Society Handbook of Physiology*, Section 12. Exercise regulation and integration of multiple systems. Oxford University Press, New York, 1996, pp. 1075–1123.
41. Baldwin, K.M., and F. Haddad. Plasticity in skeletal, cardiac, and smooth muscle: Effects of different activity and inactivity paradigms on myosin heavy chain gene expression in striated muscle. *J. Appl. Physiol.* 90: 345–357 (2001).
42. Roy, R.R., D.L. Hutchison, J.A. Hodgson, and V.R. Edgerton. EMG amplitude patterns in rat soleus and medial gastrocnemius following seven days of hindlimb suspension. *IEEE Eng. Med. Biol. (Proc)* 10: 1710–1711 (1988).
43. Riley, D.A., G.R. Slocum, J.L.W. Bain, F.R. Sedlak, T.E. Sowa, and J.W. Mellender. Rat hindlimb unloading: Soleus histochemistry, ultrastructure, and electromyography. *J. Appl. Physiol.* 69: 58–66 (1990).
44. Blewett, C., and G.C.B. Elder. Quantitative EMG analysis in soleus and plantaris during hindlimb suspension and recovery. *J. Appl. Physiol.* 74: 2057–2066 (1993).

45. Clement, G., and C. Andre-Deshays. Motor activity and visually induced postural reactions during two-g and zero-g phases of parabolic flight. *Neurosci. Lett.* 79: 113–116 (1987).
46. Roy, R.R., J.A. Hodgson, J. Aragon, M.K. Day, I.B. Koslovskaya, and V.R. Edgerton. Recruitment of the Rhesus soleus and medial gastrocnemius before, during and after spaceflight. *J. Grav. Physiol.* 3: 11–16 (1996).
47. Hodgson, J.A., S.C. Bodine-Fowler, R.R. Roy, R.D. de Leon, C.P. de Guzman, I. Koslovskaya, M. Sirota, and V.R. Edgerton. Changes in recruitment of Rhesus soleus and gastrocnemius muscles following a 14 day spaceflight. *Physiologist* 34 (Suppl.): S102–S103 (1991).
48. Recktenwald, M.R., J.A. Hodgson, R.R. Roy, S. Riazanski, G.E. McCall, I. Koslovskaya, D.A. Washburn, J.W. Fanton, and V.R. Edgerton. Effects of spaceflight on Rhesus quadrupedal locomotion after return to 1 G. *J. Neurophysiol.* 81: 2451–2463 (1999).
49. Booth, F.W., and C.R. Kirby. Changes in skeletal muscle gene expression consequent to altered weight bearing. *Am. J. Physiol.* 262: R329–R332 (1992).
50. Balon, T.A., T.E. Zirkel, B.J. Musser, C.S. Stump, C.M. Tipton, and W.L. Lowe. Impairment of insulin-like growth factor (IGF-1) gene expression by hindlimb suspension. *Med. Sci. Sport. Exercise* 25: S4 (1992).
51. Adams, G.R., and F. Haddad. The relationship among IGF-1, DNA content, and protein accumulation during skeletal muscle hypertrophy. *J. Appl. Physiol.* 81: 2509–2516 (1997).
52. Roy, R.R., C. Tri, E.J. Grossman, R.J. Talmadge, R.E. Grindeland, V.R. Mukku, and V.R. Edgerton. IGF-I, growth hormone and/or exercise effects on non-weight-bearing soleus of hypophysectomized rats. *J. Appl. Physiol.* 81: 302–311 (1996).
53. Grindeland, R.E., R.R. Roy, V.R. Edgerton, E.J. Grossman, V.R. Mukku, B. Jiang, D.J. Pierotti, and I. Rudolph. Interactive effects of growth hormone and exercise on muscle mass in suspended rats. *Am. J. Physiol.* 267: R316–R322 (1994).
54. Aboudar, S.B., B. Sempare, H. Koubi, H. Dechaud, and D. Desplanches. Effects of adrenalectomy or RU-486 on rat muscle fibers during hindlimb suspension. *J. Appl. Physiol.* 75: 2767–2773, (1993).
55. Hickson, R.C., S.M. Czerwinski, and L.E. Wegrzyn. Glutamine prevents down regulation of myosin heavy chain synthesis and muscle atrophy from glucocorticoids. *Am. J. Physiol.* 268: E730–E734 (1995).
56. Tischler, M.E., M. Slentz, A. Aannestad, J. Farah, and R. Siman. Slowing atrophy of unweighted soleus using protease inhibitors. *ASGSB Bull.* 9: 33 (1995).
57. Swoap, S.J., F. Haddad, P. Bodell, and K.M. Baldwin. Control of beta myosin heavy chain expression in systemic hypertension and caloric restriction in the rat heart. *Am. J. Physiol.* 269: C1025–C1033 (1995).
58. Diffie, G.M., F. Haddad, R.E. Herrick, and K.M. Baldwin. Control of myosin heavy chain expression: Interaction of hypothyroidism and hindlimb suspension. *Am. J. Physiol.* 261: C1099–C1106 (1991).
59. Burke, R.E., and V.R. Edgerton. Motor unit properties and selective involvement in movement. In J. Wilmore and J. Keogh (eds), *Exercise and Sport Science Reviews*, Academic Press, New York, 1975, pp. 31–81.
60. Burke, R.E. Motor units: Anatomy, physiology and functional organization. In J.M. Brookhart and V.B. Mountcastle (eds), *Handbook of Physiology. The Nervous System. Motor Control*, American Physiological Society, Bethesda, MO, 1981, pp. 345–422.
61. Saltin, B., and P.D. Gollnick. Skeletal muscle adaptability: Significance for metabolism and performance. In L.D. Peachey (ed.), *Handbook of Physiology. Skeletal Muscle*, Sect. 10, Chap. 19, American Physiological Society, Bethesda, MD, 1983, pp. 555–631.
62. Gardiner, P.F., B. Hibl, D.R. Simpson, R.R. Roy, and V.R. Edgerton. Effects of a mild weight-lifting program on the progress of glucocorticoid-induced atrophy in rat hindlimb muscles. *Pflugers Arch.* 385: 147–153 (1980).

63. Diamond, J.M. The evolution of quantitative biological design. In E.R. Weibel, C.R. Taylor, and L.C. Bolis (eds), *Principles of Animal Design*, Cambridge University Press, Cambridge, 1998, pp. 21–27.
64. Hammond, K.A., and J.M. Diamond. Maximal sustained energy budgets in humans and animals. *Nature* 386: 457–462 (1997).

KENNETH M. BALDWIN
Department of Physiology and Biophysics
University of California
Irvine, California

V. REGGIE EDGERTON
Department of Physiological Science
University of California
Los Angeles, California

ROLAND R. ROY
Brain Research Institute
University of California
Los Angeles, California

N

NASA MISSION OPERATION CONTROL CENTER AT JOHNSON SPACE CENTER

Background

As the Space Task Group (STG) began to relocate to Houston, Texas, from Langley Field, Virginia, in 1962, the last of the Mercury flights were still being supported from the Mercury Control Center (MCC) at the Cape Canaveral Air Force Station in Florida. In those early days, dramatic change was underway throughout the manned space program. Galvanized by President Kennedy's Moon challenge in 1961, Gemini and Apollo began to take form in the engineering, design, and manufacturing organizations throughout the country. The learning curves were very steep in all of the disciplines. This was equally true in the control of flight operations, as they were facilitated in the first control center in Florida and at key remote sites around the globe.

The flight operations function (becoming known as flight control) faced new changes for the upcoming programs—missions were much more complex due to rendezvous, extra vehicular activity (EVA), and lunar challenges. Telemetry and command systems were changing from analog to digital, reliable global communications were enabling a more centralized control center with fewer remote sites, and the schedule demanded extensive training with concurrent training for one flight while actively performing another simultaneously. All of these considerations and others led to the conclusion that the new Mission Control Center must be colocated with the flight control team in Houston.

Beginnings—Operating Concept

Discussion of flight control in the MCC begins with an operating concept of the way missions will be planned and managed by the team. Flight control was, and

still is, part of the overall planning effort. In general, the program office defined the objectives and selected details for each of the flights. The definition by the program office evolved over the years from rather simple statements of objectives to more formal documents, outlining objectives, flight content, and some of the sequencing of events. The flight control organization has always seen itself as responding to these program office requirements, while participating in their development. In addition to the flight control element, another analytical organization has always existed within the overall operations organization. In general, its role was the mission planning task, again in response to program objectives. It developed the software analysis tools, the options, and the recommendations in areas such as launch phase design, abort mode definition, rendezvous and entry techniques, lunar trajectory simulations, lunar navigation methods, use of propulsion systems and guidance for lunar landing, and some consumables analysis. To this overall process, the flight control team brought a detailed knowledge of how to use all of the spacecraft systems to accomplish these objectives and techniques and how to monitor the basic health of the spacecraft systems as they are performing the assigned tasks. Another part of the flight control team, specialists also in trajectory, guidance, and abort planning, took the techniques defined by the planning organization and determined how to display, monitor, and control all of the necessary real-time steps in performing these techniques.

Early in Mercury, the role and authority of the flight control team in the MCC was established and integrated with the program office, the flight crew, and the vehicle engineering organizations. That is stated rather simply now but was a major learning experience at the time for all participants. Resolution was greatly enhanced by the early leaders of the STG, some of whom had considerable aircraft flight-test experience and reputations.

During Mercury, there were also arrangements with STG and other engineering organizations to participate as flight controllers. This worked better in some cases than others, but the eventual demands of time and travel for simulations, launch site test support, and real-time flight control led to the decision to use dedicated personnel assigned full time to the flight control organization.

The flight control team in Mercury had another component that gradually reduced in size to zero as the 1960s progressed. Figure 1 is a useful reference to explain how this component came to exist.

The figure displays the ground tracks resulting from consecutive spacecraft orbits around the globe. The figure also portrays the telemetry coverage of the spacecraft available from various locations which were referred to as remote sites. These sites were in contact with the spacecraft for up to 5 minutes but real-time data was not quickly or reliably routed back to the MCC in Florida. Crew voice was generally relayed to the MCC in real time, but those communication links were also not as reliable as desired. Therefore, small flight control teams of three people were sent on station at key remote sites. They conducted air-to-ground communications with the flight crew and evaluated spacecraft systems in real time with their local displays—much like a short-term, surrogate control center. They also sent teletype snapshots of spacecraft systems data back to MCC after the spacecraft passed over their site. This worked well enough but required a significant number of people. The coordination between MCC and the multiple



Figure 1. A reconstruction of the Mercury Control Center at Kennedy Space Center shows the world map with the ground tracks and station coverage (photo courtesy of NASA #KSC-62PC-128). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

remote site teams also created the necessity to keep many people well informed and synchronized in mission awareness.

This component (active flight control teams at remote sites) was carried over into Gemini, the successor flight program to Mercury. But as global communications improved, the number of teams deployed was gradually reduced. There was an average of 13 remote flight control teams for Mercury. Seven teams with six flight controllers were still deployed early in Gemini but gradually phased out after the unmanned LM-1 flight in early 1968. Even as it phased out, the experience gained by so many young engineers, two to three years out of college, by leading one of these teams was heady stuff and accelerated their learning and growth dramatically.

The guiding principle of the Mission Control Center (MCC), flight control team was, and is today, straightforward. (Note: After the Mercury missions were completed, the term MCC was changed to mean Mission Control Center.) The team was responsible for making any decision or taking any action to conduct the mission safely and successfully. This responsibility carried the requirement to make any necessary decisions within seconds. This criterion and the temperament to accept this responsibility became the driver for selecting people. The required understanding of spacecraft systems and trajectory techniques, as well as understanding how to use all of these in varying mission circumstances was also very important. This real-time response requirement drove the need for

training and simulations and the need to establish documented mission rules and procedures by which to guide real-time decision making. Spacecraft systems and trajectory/guidance disciplines were the core skills of this team. Systems team members developed and published the spacecraft system handbooks which were functional schematics of the spacecraft systems, including the placement of telemetry measurements. Not only was this a powerful learning tool, it also became the reference for all of the operations team; they were also used extensively by the engineering and industry teams. Spurred by the false heat shield deploy telemetry indication on John Glenn's MA-6 flight, the systems controllers performed a very valuable program service in critically assessing and improving spacecraft telemetry instrumentation on all future spacecraft.

Another element of the operating concept was the focus on training. This was and is a phased process, starting in the classroom and proceeding through individual systems training in special facilities and interactive computer-based training. The final step in flight preparation involved closed-loop simulations with the entire flight control team in MCC and the flight crews in simulators. This final step tested and verified all of the planning and mission rules for each of the major phases of the mission. Later on, members of the flight control team also observed geology training for the crews, EVA training in the large water tank, and even flew jump seat in the Shuttle training aircraft as an adjunct to their training.

All of these and other considerations led to the definition of the flight control team structure. The leader and final decision maker was the Flight Director, and the rest of the team reported to this position. An astronaut was always assigned as the voice interface between the flight crew and the MCC. The designation was Cap Com from the original Mercury terminology—Capsule Communicator—and represented the crew's point of view to the MCC team. The rest of the team was populated by specialists who had console call signs such as

- *GNC—control and propulsion systems (command service module)
- *EECOM—environmental and electrical (command service module) (com was reassigned)
- *INCO—spacecraft/ground communications
- *CONTROL—control and propulsion systems (lunar module)
- *TELMU—environmental and electrical (lunar module)
- *AFD—assistant to the flight director
- *PROCEDURES—coordination with remote sites
- *FIDO—trajectory monitoring and control
- *GUIDO—onboard guidance systems
- *RETRO—abort and return to earth planning
- *BOOSTER—launch vehicle systems
- *FAO—flight planning and crew checklist
- *SURGEON—flight surgeon
- *NETWORK—interface with the network
- *M & O—interface with MCC systems

Each one of these specialists was supported by additional team members located in staff support rooms (SSRs), often called “back rooms.” The number of SSR participants varied for each discipline and by the phase of the mission. This feature provided in-depth support to the Mission Operations Control Room (MOCR) console operators and an extra resource to pursue resolutions of multiple problems as they occurred. The full MOCR teams numbered about 16 per shift. SSR support averaged about 106 personnel on each shift. Besides this core team, there were several other consoles for management, public affairs, and other involved agencies. The flight control team also had the responsibility to define all of its requirements for data displays, commands, and voice loops for each of the respective positions in MCC, preparing them to use these capabilities.

During the preparation time leading up to a flight, the flight control team engaged the program office to ensure understanding of objectives and to influence specific details. Very importantly, controllers were also active participants in the planning process, helping to define the specifics of the mission, flight plan, checklists, and their own procedures. The Flight Director and his flight control team also updated the mission rules for the unique aspects of a particular mission after detailed consultation with the flight crew.

To illustrate the relationships between the various documents, the mission plan was the source material for trajectory design, usually in the form of multiple reports. The flight plan was a detailed time line of flight activities prepared for the flight crews to use but monitored and followed by all parties. Mission rules governed the criteria for deviating from the nominal plan and selecting alternative courses of action up to and including early mission termination. Checklists and procedures were also referenced in the flight plan and governed detailed switch by switch scenarios for standard activities such as guidance platform alignments and EVA suit donning. Spacecraft malfunction procedures were developed by astronauts and flight controllers. Procedures were also documented for each console operator, outlining their detailed actions for using the capabilities at their consoles. Systems handbooks were references for troubleshooting.

The flight control team defined the interface with the launch site team, the range safety office, and any other mission support functions. These other functions might include payload centers operating within the MCC or at remote locations such as Goddard Space Flight Center, (GSFC), Jet Propulsion Laboratory (JPL), Marshall Space Flight Center (MSFC), and the Air Force Space Flight Control Center at Sunnyvale, California (only for the Shuttle). This interface definition included requirements for voice and data transfers between these locations and the procedures governing any transactions. Procedures also included hand-over definitions such as the transfer of control from the launch site to the MCC at clearance of the launch tower. The flight control team then conducted simulations with these other organizations to prepare for the flight. At the end of this process, the teams performed their flight support roles during the mission.

Although the type and number of experts varied with time or specific mission requirements, this operating concept endures until today. It has served the programs and NASA very well during the past decades and will continue to do so.

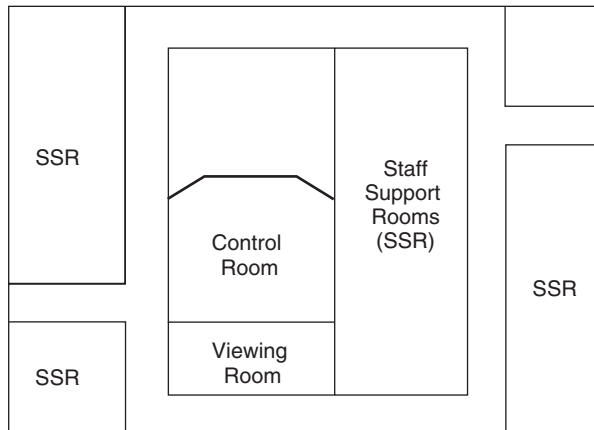
The Control Center is the instrument by which this team performs its task. Based on the Mercury experience and the vision of what Gemini and Apollo

would require, the Mission Control Center (MCC) in Houston began to be defined starting in 1962. This definition recognized that the MCC would evolve in equipment technology and the scope of participation. The MCC was located in Bldg. 30 at the Manned Spacecraft Center (MSC, later renamed as the Johnson Space Center—JSC) complex. Building 30 included an office wing and the Mission Control Center element. The MCC wing was a three-tier facility of 118,500 square feet; the computing and communications equipment are on the first floor, topped by two floors of very similar design. Each has a central Mission Operations Control Room (MOCR), later called Flight Control Room (FCR) for the Shuttle, a viewing room behind the MOCR, and multiple staff support rooms (SSR—later called MPSR's, multipurpose support rooms for the Shuttle) which provided in-depth support to the flight controllers in the MOCR. Additional space on the operational floors was assigned to the simulation team, the recovery team, and other support functions. Figure 2 is an overview of the MCC layout.

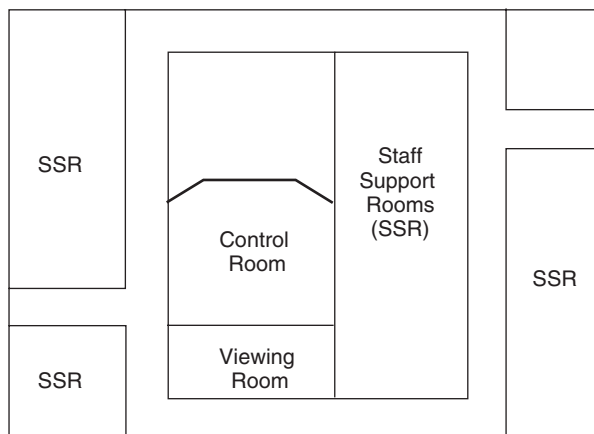
Figure 3 is a layout of the MOCR that has consoles for all of the discipline specialists in the room. The MCC in Houston was implemented under a 1962 contract with Philco-Ford Western Development Laboratories (later moved to Houston as Philco Houston Operations—PHO) for the design, development, and maintenance operations of the overall facility. NASA also contracted with IBM in 1962 for the Real-Time Computer Complex (RTCC). These contracts were modified, extended, and/or recompeted until a major change occurred in 1986 (to be discussed later).

As the concept of MCC as the central hub of ground based flight control continued to evolve for Apollo, floor space and equipment was dedicated to representatives of the spacecraft program office who orchestrated support to each mission from JSC engineering and vehicle contractors. The role for flight control remained the same. All decisions and actions were taken by this team especially for near real-time decisions, that is, within seconds or minutes. When an unanticipated technical condition occurred and “time-to-decide” was measured in hours or days, there was an opportunity to consult the program office/engineering/industry team for help in the resolution. This arrangement was not very strongly implemented for Gemini, perhaps because the program office did not have a test and evaluation organization as in Apollo and also because much of the JSC engineering was focused on Apollo and not Gemini. For Apollo, this concept greatly added to the strength and problem-solving capabilities of the MCC team. The spacecraft program office support function (including flight control and spacecraft industry representatives) came to be known as SPAN (Spacecraft Analysis room). This room was tied into other facilities outside the MCC that housed JSC and industry engineering teams. These teams followed the missions with telemetry and voice available from all of the key communication loops. This interface (SPAN) tied the rest of the program team into the MCC in a very controlled and orderly way. The arrangement ensured high confidence in dealing with anomalous technical conditions and a unified, well-understood methodology for engaging the program office team of JSC engineering and vehicle contractors.

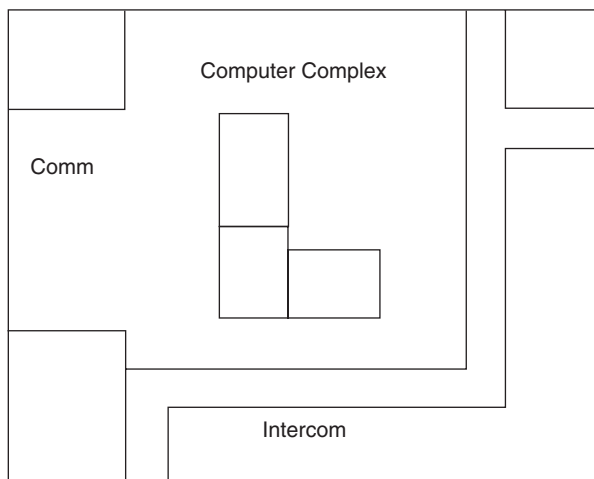
Additionally the MCC provided a facility for the interface with the payload customers, principal investigators, and scientific teams that sponsored individual payloads. Again, this methodology ensured the highest confidence in dealing with any deviations to the planned payload time line or use of that equipment.



Third Floor



Second Floor



First Floor

Figure 2. Building 30 MCC at the Johnson Space Center has three floors; the upper two floors are nearly identical.

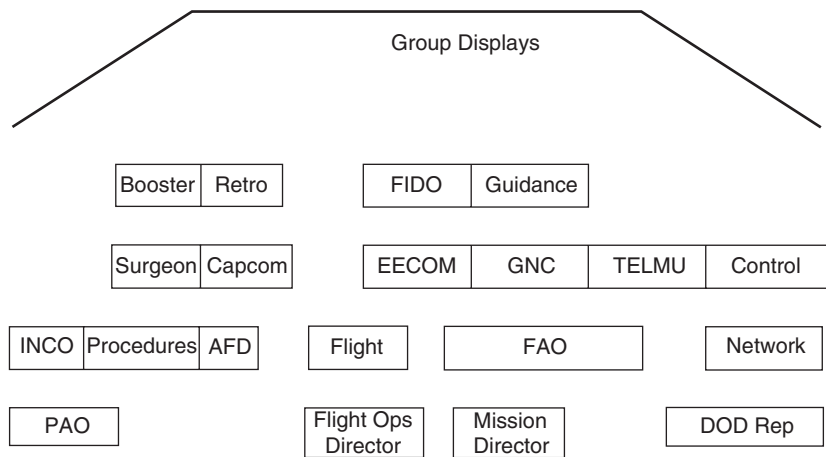


Figure 3. The MOCR (FCR) layout at the Johnson Space Center included the Flight Control Team in the front three rows and a back row for management and administration.

Figure 4 is an illustration of MCC as the central hub, tying in various organizations for decision making on ongoing flight activity.

Gemini Era—1962–1966 (1)

In the Gemini program, the mission planners, the flight control teams in MCC, and the flight crews established the operational experience base and confidence for Apollo. Gemini was conceived and conducted to test as many of the Apollo

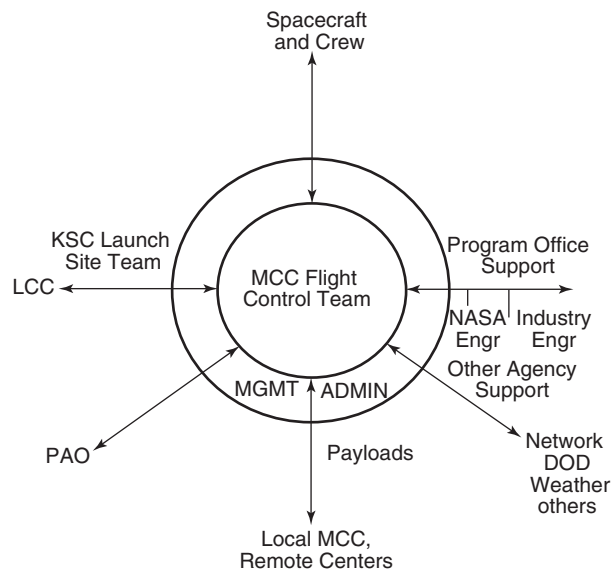


Figure 4. The MCC Flight Control Team is the hub of all ground support functions during missions.

mission techniques as possible in Earth orbit. As the reliability of global communications grew, the Gemini operations began to reduce the number of remote-site, flight control teams and centralize decision making in the new MCC. Gemini was a fast moving program that had multiple technical challenges. Ten manned Gemini flights were planned and flown in a 20-month period. From the first short-duration flight of Gemini 3, Gemini 4 accomplished an EVA and experienced the first difficulty in attempting to fly formation, using the last stage of the Titan launch vehicle. A long duration was the emphasis of Gemini 5 for almost 8 days. After an Agena failure, two Gemini spacecraft on 7/6 completed the first rendezvous and station keeping and achieved a nearly 14-day flight by the Gemini 7 crew. Gemini 8 docked with an Agena stage, but a short in the Gemini control system led to an emergency undocking and early mission termination in the Pacific contingency recovery area. Gemini 9–12 experimented with multiple types of rendezvous, docked, and tethered Agena maneuvers. These final flights also concluded the learning process for orbital EVAs. Once adequate crew positioning aids (hand- and footholds) and water tank training were made available in Gemini 12 spacecraft, the crews had a much more controlled way to perform assigned EVA tasks. With all of these new challenges, many missions also experienced spacecraft problems especially with the fuel cells and control systems thrusters. All of this provided even more experience in real-time problem solving, while continuing to satisfy mission objectives.

The software capability to handle these new operational challenges was embedded in the MCC systems (2). These systems can be technically described around the three major functions of MCC architecture: the Communications Interface System (CIS), the Real Time Computer Complex (RTCC), and the Display and Control System (DCS). The CIS provides internal communications and connects the MCC to the spacecraft communications links via the network and remote sites. The RTCC provides the mainframe computational and display processing, and the DCS provides the information to the human operators.

The communications system elements consisted of a UNIVAC 490 (the communications processor), a Lockheed PCM-102 ground station, and some custom hardware, called the Master Digital Command System (MDCS). The Gemini launch data systems collected data on 2.0-kbps lines from the Eastern Test Range (ETR), as well as on 40.8-kbps lines where the ground station did the frame-sync and decommutation functions. Each console had an intercom panel for the major loops such as air to ground, flight director, and launch side coordination and a set of special loops unique to each console.

The mainframe processors consisted of five IBM 7094s, using a customized IBM operating system within the RTCC. The machine which was designated as the mission operations computer (MOC) software would process the data from the communications processor and perform the telemetry, trajectory, and command functions. A concept of a dynamic standby computer (DSC) was established. The DSC processed the same inputs with the same software, but the results were not used. The computer operators would compare the results from the two machines to ensure that they were synchronized. Should the MOC fail, the DSC would be brought on line within seconds to become a new MOC.

The DCS element that created video displays and presented them to the operator was the digital to television (DTV) subsystem. The MOC data was

shipped to the DTV subsystem across a channel buffer. The DTV subsystem then used the MOC data to produce 48 channels of display. The background format consisted of fixed display formats which were determined premission and did not change with telemetry processing. The foreground data elements were the real data resulting from real-time telemetry. This system provided numerical outputs on black and white CRTs with essentially no graphics. Figure 5 is a typical Gemini era console.

The projection plotter display equipment was the subsystem that produced the large screen display of the spacecraft position and ground track on the world map, as well as other launch and landing graphics (2). Other MCC-unique systems were developed in this time frame, including the video hard copy system. If an operator needed a printout of his video display he pressed a hard copy request button which activated a camera that took a 35-mm film image of the same video channel that the operator was viewing. The film was automatically processed, printed on paper, and dried in fifteen seconds. The hard copy equipment operator then placed the print into the pneumatic tube (P tube) system, and the canister delivered the processed print to the console operator (2).

In the original Mercury Control Center, the computing center was managed by a team from the NASA-Langley Research Center destined to join what became the Goddard Space Flight Center (GSFC) in Greenbelt, Maryland. Early in the software development, the computing facility was actually housed in an IBM building on Pennsylvania Avenue in Washington, D.C., but the actual flight support to the Mercury Control Center was in a computer complex at GSFC. Its role was limited to launch tracking, orbit determination (navigation), retrofire,



Figure 5. A typical console for Gemini and Apollo featured two CRT displays, intercom panel, event and status lights, and command switches (photo courtesy of NASA; JSC-#68-49299). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

and landing calculations because it did not process telemetry. Telemetry and command functions were processed through separate equipment in the Mercury Control Center in Florida and the remote sites. Because this computing function was so vital, new, and so geographically separated, a backup capability for retrofire calculations and other trajectory support was established in a facility called the Auxiliary Computing Room (ACR). The ACR employed the same programs used for premission planning and the institutional computers at STG in Virginia. This backup capability was carried over into the new MCC in Houston using institutional mainframe machines in the JSC complex. During Gemini, additional premission planning programs in the ACR were used to back up rendezvous maneuver calculations and entry planning. This capability was eventually phased out after Apollo 12.

In Gemini, the simulation task became much more sophisticated than in Mercury. Mercury basically used an open-loop simulation system, whereas Gemini had to be closed loop between MCC and the Gemini simulator because of more variations in potential actions by the crew or the MCC. The Gemini simulator became the major training platform for the flight crews. The simulations were mechanized and conducted in closed-loop fashion, so that all actions at either the simulator or the MCC were correctly reflected in the ongoing simulation. A simulation control area was located on the second floor of the MCC, where the simulation team managed the complex process of establishing realistic and challenging exercises. During this period, some flight control teams were still sent to selected remote sites before adequate data communications were available. The simulations area also had a replicate of the remote-site control rooms for training the remote-site teams.

In March 1965, the MCC-H monitored the Gemini 3 mission, and primary control responsibility rested with the old Mercury Control Center at Cape Canaveral. Although not all systems were completely tested and operational, the MCC-H came on line in such a manner as to convince NASA that it was ready to control Gemini 4. In June 1965, NASA manned the new MCC in Houston to control the Gemini 4 mission with MCC in Florida as backup. In the flight of Gemini 4 that included the first space walk by an American astronaut, the MCC in Houston was declared operational and the old control center at Cape Canaveral was deactivated. The H was dropped from the name and Houston became the control center for all subsequent manned spaceflights (2).

The Gemini experience base in operations from the MCC provided a powerful level of confidence as the planners, flight control teams, and flight crews moved to the Apollo program.

Apollo Era—1966–1972 (1)

While Gemini was underway, unmanned Apollo flights were scheduled to test the new launch vehicle (Saturn 1B) and the Apollo command and service module (CSM). During 1966, two successful flights (AS 201 and AS 202) were conducted with high-speed entries to test the CSM heat protection system. These flights also provided an opportunity for the flight control team to understand the Apollo spacecraft. After the last Gemini flight in November 1966, Apollo was ready to move into the manned flight phase.

Then, in January 1967, tragedy struck Apollo with the fire on the launch pad. Three crew members perished. The program went into an agonizing reappraisal. Externally, some questioned whether the program should continue. But internally, the recovery process and spacecraft modifications moved into high gear. The country and Congress reaffirmed their support for Apollo. The unmanned flights resumed with the first two flights of the Saturn V and reentry of the CSM at speeds equal to lunar return velocities. The second Saturn flight produced a major scare when two engines shut down late in the second stage due to a center engine pogo condition and a miswiring of the shutdown circuitry. But it did not significantly affect the mission. An unmanned lunar module (LM) was also successfully tested in low Earth orbit after it was launched on a Saturn IB. These unmanned flights added specific Apollo knowledge to the Gemini experience of the flight control team.

This operating capability enabled the Apollo program to fly 10 days on the first manned orbital mission (Apollo 7) and to go to lunar orbit on the second manned Apollo mission (Apollo 8). Two more major missions, Apollo 9 and Apollo 10, tested the CSM and LM in Earth orbit and then in lunar orbit. Apollo 11, the fifth manned mission in the program, landed on the moon and returned safely. During the powered descent of the lunar module (LM), the flight control team evaluated computer program alarms, and correctly assessed them as acceptable to proceed. This schedule represented a very aggressive sequence of missions through Apollo 11, which would have been unlikely without the Gemini operating foundation. Apollo 12 was struck by lightning soon after liftoff and still reached a nominal Earth orbit, but the CSM guidance platform was lost; the CSM guidance was restored in orbit. MCC evaluation of the CSM and the Saturn IV B permitted a "GO" for translunar injection to the Moon. With the aid of MCC lunar navigation, Apollo 12 was vectored so that the onboard LM systems and the crew guided the landing to the surveyor spacecraft, which had landed on the Moon years earlier. Subsequent precision landings allowed specific planning and execution of scientific exploration of the lunar surface on five landings after Apollo 11. The successful Apollo 13 return was the ultimate test and demonstration of capability by the flight control team, ably assisted by the engineering/industry teams around the country, for problem resolution with a long "time to decide." All of the CSM power-up and entry procedures were tested in the simulators by astronauts. This success was the result of a decade-long improvement in the flight control team in solving complicated problems in real time and the on-call engineering support provided by the program office team. Apollo 14 was the recovery flight after the Apollo 13 mission. Apollos 15, 16, and 17 extended lunar surface exploration with longer lunar stays, more EVAs, and the lunar rover for surface mobility. The CSM also had a service module bay full of scientific instruments for mapping the Moon and its environment while in lunar orbit. Apollo was also the first time the flight control team enjoyed the luxury of full-time coverage from the Deep Space Network (DSN) once the vehicle left low Earth orbit and while it remained on the front side of the Moon.

Any discussion of the MCC system would be incomplete without recognition of the management processes involved. Because the mainframe computers used in the MCC were always fully taken up with the telemetry, command, and trajectory planning requirements, there was very high priority in establishing a

highly reliable MCC system. This introduced very real tension between the flight controller's desire for new and improved capabilities (requirements) and the implementer's desire to develop and freeze a system (implementation) which would be stable and reliable for each flight. This tension resulted in major and painful "scrubs" of requirements to fit within the system capacity and the program schedule. In spite of the pain, the process worked well due to the guidance and dedication of the flight organization management team. Their skills generally resulted in the most capability available for the flight control team at any given time, and maintained rigorous control of the implementation process and the delivery of a reliable system.

To elaborate further on the software implementation, it was not an art, as some considered software development at that time, but a well-practiced discipline. The management process consisted of three basic elements: planning, execution, and feedback. The planning phase for the large software system development consisted of the following steps: (1) define the goals and objectives, assign responsibilities, and organize; (2) define the system requirements; (3) design the system; and (4) estimate the required resources. The execution phase was divided into the following steps: (1) implementation or programming, (2) testing, (3) verification, (4) operation, and (5) maintenance. The feedback phase consisted of reviewing progress, incorporating changes, and updating plans. These three phases were defined somewhat arbitrarily for presentation and, in most cases, were not sequential. However, all of the steps did exist and had to be reckoned with, especially because the software had to accommodate changes reasonably (3).

There were no cases where MCC did not have adequate capability for each flight or where the MCC was not ready to support. Sometimes, the desired stability was reached uncomfortably close to the scheduled time for the system's use. But in the end, MCC was always there and ready. This performance was a major tribute to the managers of the both requirements and implementation processes.

During the Apollo project, the CIS communication processors were upgraded from Univac 490s to 494s. The extra processing power allowed the 494s to perform some of the PCM ground station functions and all of the previous MDCS functions. The new merged system was called the Communications Command and Telemetry Subsystem (CCATS). The CCATS in the MCC provided the capability of simultaneously performing the functions of digital communication, data handling, telemetry data, decommutation and distribution, and verification and control.

Similarly, the RTCC processing systems outgrew the IBM 7094s, and they were replaced by five 360/75s. Some of these were the first units off the IBM production line. The concept of using a dynamic standby computer which shadowed every operation of the MOC was carried over from the Gemini design. The operating system of the 360s was modified to support the performance needs of the MCC, and the new operating system was named The Real Time Operating System (RTOS), later upgraded to the Extended Operating System (EOS) (2).

The increasing performance needs were primarily due to the increased complexity of the trajectory functions of a lunar mission, but there were also new changes in the type of data which was being processed. There was an increased emphasis on flight planning aids and decision-making support.

Looking back on those days, it is very interesting to note that the MCC supported the lunar landing missions with a total computer capacity comparable to one modern work station and the new MCC has hundreds of these work stations.

As part of the continuing provision for payload support, a data system was developed in the MCC to support the Apollo Lunar Surface Experiment Package (ALSEP); the first of these was deployed on Apollo 12 (2). A simpler, solar-powered version was deployed on Apollo 11. These packages had a complex of scientific instruments, including seismic detectors, and were designed to remain on the surface and take measurements for years. An operational control room was implemented on the MCC second floor using standard consoles and displays for this package. The ALSEP data system was deactivated in September 1977 after approximately 8 years of continuous and successful support of the five operational ALSEP packages left on the Moon on each landing mission (2).

Skylab, Apollo-Soyuz Test Program (ASTP)—1973–1975 (1)

Skylab, the first U.S. space station, was designed to conduct solar, Earth resource, biomedical, and a subset of corollary experiments. It was also the platform for long-duration in-orbit experience for U.S. astronauts. The initial unmanned launch of the Skylab station in May 1973 resulted in the loss of one of two solar wings and the meteorite shielding during the launch phase. The vehicle came close to becoming a complete failure once it reached orbit because of the increasing temperature within the living module caused by the loss of the shielding. These conditions were repaired by the first Skylab crew, using a manually deployed thermal shield (like a parasol) through one of the Skylab's airlocks from inside the module. An EVA was used to restore full power from the remaining solar wing. On the second flight, the crews replaced the initial thermal shield with a more permanent design, and the third manned mission culminated in an 84-day record flight. In-orbit operations were a real test of the flight crews, the flight control team, the scientific experiment teams, and the MCC itself. The self-imposed demands for maximum return on the time spent in Skylab caused some amount of work overload on the crew and on the MCC operators and occasionally resulted in frustration.

In July 1975, the first joint American/Soviet Earth orbital mission was conducted. In a test of rendezvous and docking compatibility, the ability of the two major space-faring nations to cooperate with each other was also tested. This experience gave the country a foundation on which later decisions were made in the 1990s to cooperate with Russia on the International Space Station. ASTP also had the communication advantage of using one high-altitude NASA communication satellite that provided coverage across half the globe.

In preparation for these missions, one of the major changes in the MCC systems was the replacement of the D/TV system with digital television equipment (DTE). A prototype DTE system was installed in July 1970. The operational DTE system supported Skylab, ASTP, and later Space Shuttle missions. Besides the DCS changes, other major modifications included implementing a data compression system on the network and the necessary processing changes in the

MCC to handle this compressed telemetry data. A major data storage subsystem was also added, called the Mass Data Storage Facility (MSDF). The MSDF consisted of two Control Data Corporations (CDC) computers, a Cyber 73 and a Cyber 74 and associated peripherals. The MCC also provided some scientific data processing in the form of an Earth Resources Preprocessing Production System (PPS). The system accepted raw Earth resource data obtained from both aircraft and Skylab before use for data application purposes. The system accepted various data formats and produced preprocessed data for distribution to Principal Investigators (PI) where final processing was performed. Skylab postmission processing continued in the MCC for years after the last Skylab mission in special scientific data systems like this one (2).

Space Shuttle—1977–1986 (1)

The Space Shuttle was the next major program in human spaceflight. National policy dictated the Space Shuttle as the single launch vehicle system for all U.S. payloads and as the carrier for any international payloads which the United States would launch. By this policy, all other U.S. launch vehicles would be phased out during the 1980s. There were also strong commitments that the Shuttle cost and price to customers would be significantly reduced from current launch vehicle systems. These two conditions—service all type of payloads and achieve major cost reductions—became significant drivers to the program. The subsequent failure to achieve expected cost reductions led to many criticisms and a lack of acceptance on the part of some of the U.S. payload community. Even without these conditions, the development of the Shuttle was a very major technical challenge. Ultimately, the challenges were met, and the Shuttle is now a major element of the U.S. launch vehicle stable.

As part of that development, there was a series of five flights of the orbiter, dropped from the Shuttle carrier aircraft at Edwards Air Force Base during the second half of 1977. These were the first Shuttle program flights, and they were supported by the flight control team. Starting in 1981, the flight control team supported the next 25 Shuttle flights, including the loss of the Challenger and crew in January 1986. Until that event, major progress was made in transiting the Shuttle system into a very versatile payload carrier and mission platform. Highlights included four orbiter flight-test missions, the launch of numerous commercial communication satellites, the first American woman in space, Space Lab scientific missions, an untethered spacewalk with a manned maneuvering unit, a U.S. senator and congressman as passengers, several international passengers, retrieval and repair (or recovery) of satellites on three separate missions, two classified DOD missions, and numerous technical advances within the Shuttle system itself. However, this trend changed abruptly with the Challenger accident in January 1986. U.S. policy was changed to restrict the use of the Shuttle to NASA and certain selected missions. The commercial community of communications satellites was not permitted to fly on the Shuttle. The DOD chose to return to expendable launch vehicles, although there were several DOD flights on the Shuttle in the early 1990s. The Shuttle payload traffic then contracted to essentially NASA and international scientific missions.

MCC supported all of these Shuttle missions. A special system was developed for the 1977 approach and landing test (ALT) flight while the total Shuttle system was still in development. The ALT system consisted of a 9.6-kbps line from Edwards to the MCC. Displays were available through the DTE system. An early version of the telemetry program was used as part of the software system. The primary operations support room was converted from the Apollo era recovery support room. Standard consoles along with X/Y plot boards were used for displays (2).

The communications interface subsystem added some major new equipment to support the Space Shuttle program. The Tracking and Data Relay Satellite Systems (TDRSS) network multiplexer-demultiplexer system came on line for STS 3 to set the stage for spacecraft communications in the TDRSS era. The first TDRSS satellite was launched on STS 6, and with later launches of TDRSS, became the system by which Shuttle data, voice, and video were communicated to the MCC (2).

Telemetry processing computers (TPC) were Interdata 832 computer systems and included the template for the current downlink telemetry format. They identified and transmitted the data across an interface to the MCC. Additionally, the onboard Payload Data Interleaver (PDI) placed payload data in "windows" in the telemetry stream, and the TPC routed the PDI traffic for delivery to local or remotely located payload customers. The TPC also drove selected analog and event displays and strip-chart recorders.

For the RTCC, the IBM 360/75s were replaced by four 370/168s. These computers performed many other functions in addition to the MOC and DSC functions. The machines supported engineering analysis, software development, vehicle trend analysis, MCC system reconfiguration, mission planning, and other functions. In the display and control system, some new custom design work on the DTE functions allowed expanding the capability. This expansion boosted the number of video displays on consoles from 80 to 104 channels (2).

Because the Shuttle was defined as the primary U.S. system for all payload communities, NASA provided special accommodations to serve their respective requirements. These accommodations were for DOD, scientific payloads, and commercial customers.

Early in the Shuttle program, the DOD identified the need to support classified flights in the MCC. In response to this requirement, implementation of a "control mode" for secure operations on the third floor was begun in 1979. Certification was approved in May 1983. Secure television was established. The mainframe computers could be switched and isolated to support classified operations. Cable trays were separated into classified and unclassified. Likewise the grounding system in the MCC was modified. Physical controls were established for separation and access. Card readers were installed, and tourist traffic was rerouted only to the second floor viewing room of the MCC.

The Payload Operations Control Center (POCC) facility provided Spacelab experimenters with a highly capable data evaluation processing and planning facility. This facility was implemented during 1981–1983 to support Space Lab flights (see Fig. 6).

Primarily for the commercial satellite customers, the general purpose Customer Support Room (CSR) was essentially a minicontrol center with

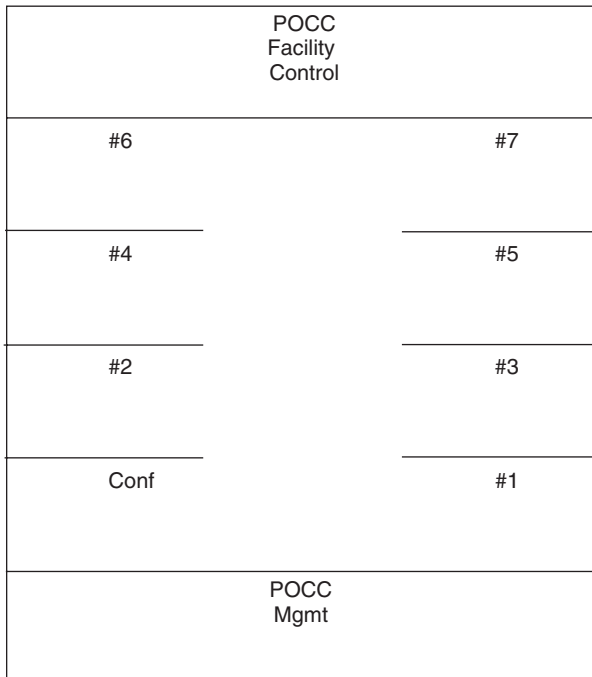


Figure 6. The Payload Operations Control Center (POCC) supported primarily the experiment teams for Space Lab missions.

communications, displays, a management meeting area, and even a viewing room (see Fig. 7).

Much of this service provided in the MCC for customers came to a halt or was redirected after the Challenger accident.

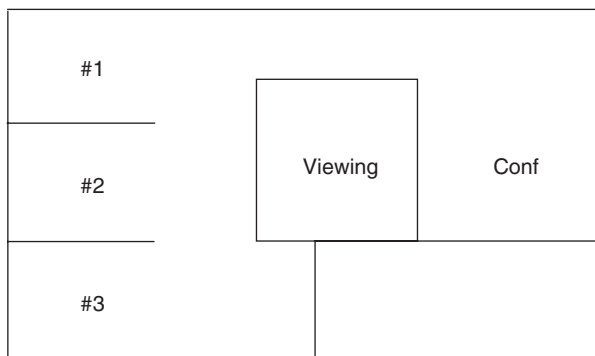


Figure 7. The Customer Support Room supported primarily the commercial community, especially the Communication Satellite Teams.

Challenger Recovery Through 1990 (1)

The recovery from the Challenger accident was again a painful process, reminiscent of the Apollo fire, for all elements of the manned spaceflight program. All of the questions asked after the Apollo fire were reengaged. Although there was an extraordinary number of detailed subjects, they were all aimed at the questions, How to recover and how to prevent another accident.

Slowly, the entire program team worked through their respective areas searching for better answers. Many answers became standard practice and resulted in more checks and constraints on the program. Safety again became the watchword for this assessment period and beyond. Eventually, the Shuttle returned to flight on STS 26 in September 1988. Then, a series of STS flights followed with significant achievements, and the flight control team contributed its share to these successes.

Shuttle flight highlights reflect the deployment of major payloads that were delayed by the accident but still scheduled for launch on Shuttle for technical or program reasons. STS 26 was the return to flight voyage of Shuttle and deployed the TDRSS 2 spacecraft. Five classified DOD missions were flown in this period. TDRSS 3 was also deployed, followed by three interplanetary missions—Magellan to map Venus, Galileo to explore Jupiter, and Ulysses to explore the polar regions of the Sun. Another Spacelab mission (Astro 1) was conducted, and STS 31 successfully deployed the Hubble Space Telescope (HST), but the optics problems encountered set the stage for a later repair mission.

In the MCC systems, the 370/168 mainframes were replaced by 3083JXs during 1986 and later updated in 1990 to 3083KXs, which doubled the CPU capacity (2). In 1986, the first real change of the major MCC contractors (PHO and IBM) since 1962 occurred when the Shuttle Transportation Systems Operations Contract (STSOC) including all MCC functions was awarded to Rockwell. During this period, two additional significant initiatives were pressed by the Mission Operations Directorate which was now the consolidated organization for planning, training, and controlling the missions, plus all the development and operations of the MCC and simulator complexes. The flight control team had been considering and experimenting with work stations in support of consoles. This was motivated by a desire to exploit new technologies and to reduce the lead time and inflexibility of the highly controlled mainframe software. This initiative was strengthened during this period by the addition of a contractor (Ford Aerospace) devoted primarily to new developments, which could be phased into the MCC, which was then separately sustained and operated by Rockwell. This accelerated the prototyping of new ideas in a test facility called the Transition Flight Control Room (TFCR), featuring workstations with color displays and graphics. The second major initiative was the early recognition of the facility needs of the emerging Space Station program. As a result of this assessment, a new wing called Building 30 South (30S) was approved, and construction began in 1990. It was planned to house the flight control team and other associated flight support activities for Space Station operations. This new wing was attached to the original MCC wing of Building 30, and provided the additional floor space deemed necessary for the task of Space Station flight control. Figure 8 portrays the five floor layout of this 103,400 sq.ft. facility.

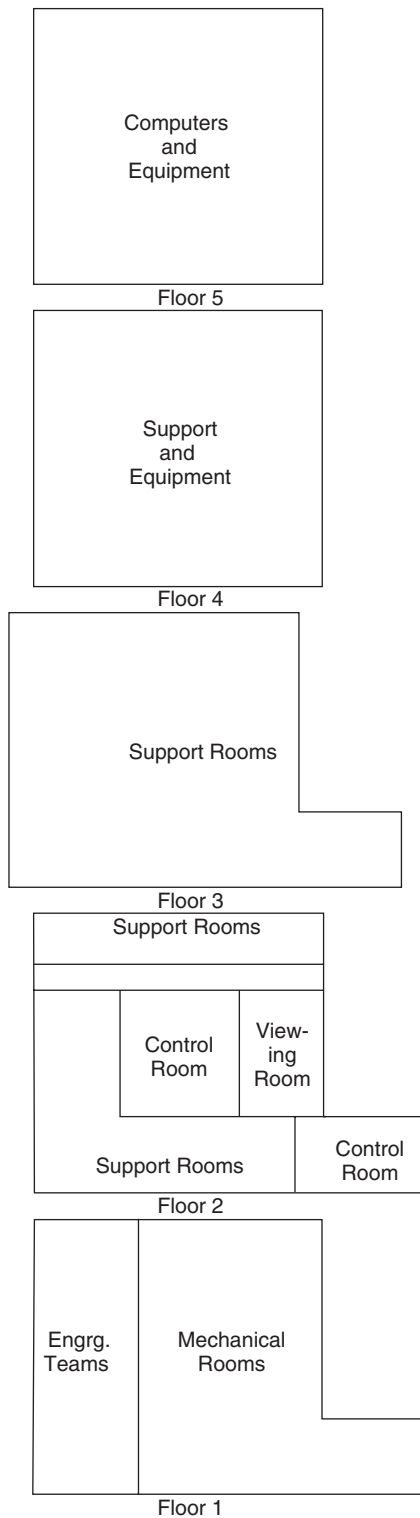


Figure 8. Building 30S MCC has five floors and two control rooms.

Shuttle and Space Station Preparations—1991–1999 (1)

For mission operations, the decade of the 1990s can be characterized by continuing support of Shuttle missions mostly with NASA payloads and planning for a newly redesigned International Space Station (ISS), including the new MCC. The 1993 decision to include the Russians led to an ISS configuration, in which the Russians would provide about 10% of the mass of the in-orbit Station complex. The substantial contribution of Russian hardware created the need for the flight control team to integrate their operations with the MCC in Moscow.

Although this was not the original intent for 30S, it became clear by 1993 that all Shuttle and Space Station mission support should be located in the single new MCC in 30S, resulting in more simplifications and efficiencies.

The Shuttle program continued to launch and service individual projects. In the 9-year period from 1991 to the end of 1999, the Shuttle program completed 58 missions at an average rate of 6.5 per year. The Shuttle flight program logged many achievements. Selected highlights of this period were three more TDRSS flights, another DOD mission, and multiple scientific missions aimed at atmospheric, life sciences, and microgravity studies, including some sponsored by other countries. The Compton Gamma Ray observatory and the Chandra X-ray telescope were also launched. There were two tether flights, a reboost of the stranded Intelsat communications satellite with the real-time decision to use three crew members (instead of the usual two) on the EVA to capture Instelsat, Hubble Telescope repair and reservice missions, and a reflight of Senator John Glenn. The Shuttle also flew one rendezvous and station-keeping mission with MIR and nine docking missions to MIR. The United States also launched its first major space station element—the node—late in 1998 and a later flight serviced the initial ISS configuration.

In 1994, the concept of a number of test Shuttle missions to MIR was established. Recognizing the huge construction task which the ISS represented and the complex coordination with Russia, it was decided to use the current Russian Space Station called MIR as a test platform to gain the operational experience of conducting joint missions and the early flight of some of the payload equipment planned for the ISS. In a sense, the Shuttle/Mir project was conceived to contribute to the ISS program what Gemini contributed to Apollo, that is, an early operational foundation.

The Shuttle/Mir project was an aggressive flight sequence over three plus years from early 1995 to mid-1998. STS 63 was the first rendezvous and station-keeping test, but without docking because the provisions were not yet available. There followed nine Shuttle docking missions, rotating crew members, and bringing supplies to and returning equipment from the MIR. Seven American astronauts lived and worked on MIR for a cumulative total of approximately 908 days, and the Shuttle brought up about 20 tons of supplies and returned 9 tons to Earth. The flight sequence continued, even in the face of major and minor MIR difficulties. The major problems were an onboard fire and depressurization of one of the MIR modules. The NASA/Russia team persevered through these difficulties and gained robust experience in jointly managing and conducting the Shuttle/MIR project.

As ISS development matured, the initial deployment of ISS hardware to Earth orbit followed. The ZARYA Russian module was launched in November 1998, followed by the American node UNITY on STS 88 in December 1998. This initial configuration was serviced on a 1999 Shuttle flight. The Russian ZVEZDA service module (containing life support and propulsion systems) was long delayed but was finally launched in July 2000. This service module was the key element, enabling continuation of ISS construction.

The design of the new MCC provided the floor layout of a Flight Control Room (FCR), as shown in Fig. 9; a typical console is also displayed in this figure. There has been a continuing interface with the MCC in Moscow since the first launch in 1995. This interface was supported by controllers within the MCC in Houston periodically and by a small team of U.S. resident flight controllers in the MCC in Moscow. Figure 10 portrays the new MCC as the central hub for the ISS, serving the requirements of both the Space Shuttle and the Space Station.

The transition to workstations in the MCC began with the early deployment of 66 workstations in 1991 that grew to more than 400 DEC Alpha 9000 workstations in operation in 2000. A high level version of the architectures of both MCCs is displayed in Fig. 11, workstations added considerable local console processing compared to the original mainframes that delivered data in fixed formats to consoles.



Figure 9. The Control Room in Building 30S layout is functionally very similar to the earlier MCC (shown here for Shuttle operations). The Flight Director's console is typical of the new MCC, and it features CRT displays with color graphics and a digital intercom control panel (photo courtesy of NASA-JSC#S95-11222). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

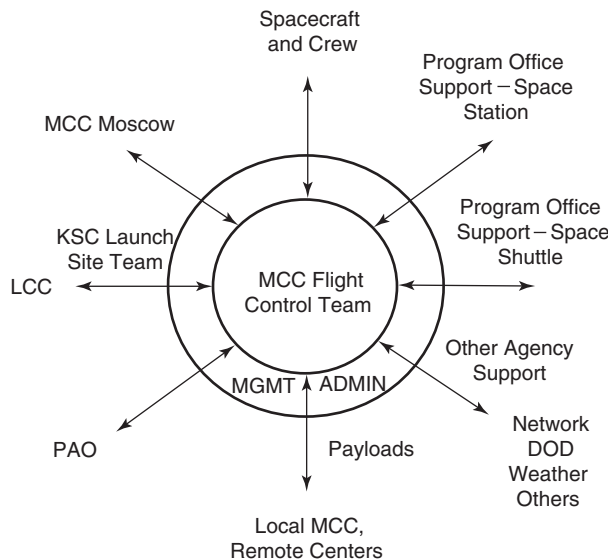


Figure 10. The MCC Flight Control Hub has new interfaces with the Space Station Program Office Team and the MCC in Moscow.

The new capability in 30S was gradually phased into operation by carefully progressing from limited flight following Shuttle flights as early as 1993 to primary use of the new MCC for in-orbit control of STS 70 in July 1995. Shuttle landing control was exercised on STS 73 later in 1995. In May 1996, all Shuttle phases were controlled from the new MCC on STS 77. (The one specific function of launch trajectory tracking and determination was still performed in a MOC on

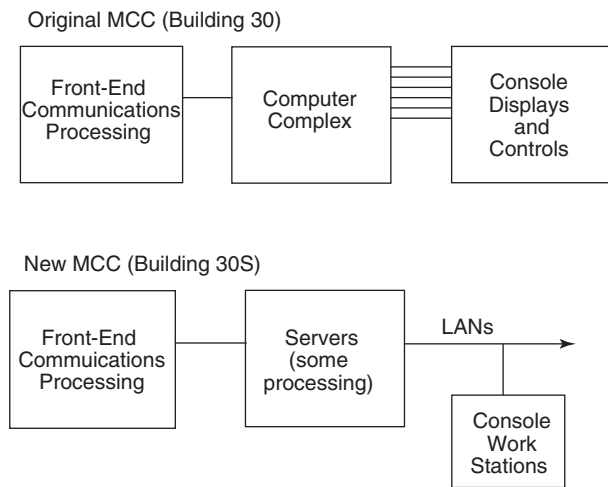


Figure 11. The evolution of MCC data architecture that now uses servers, local area networks (LANs), and workstations.

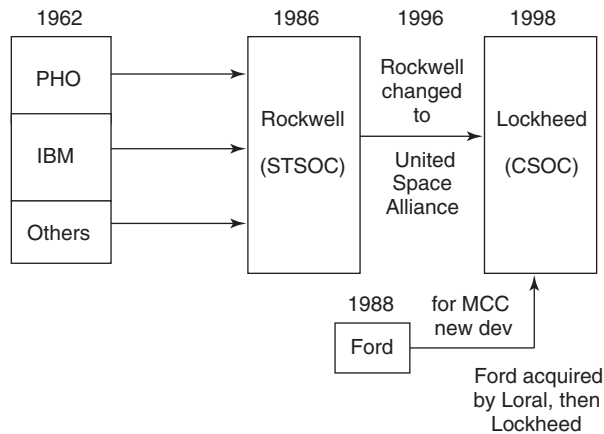


Figure 12. The history of major contractors and contracts for the MCC is shown across almost four decades.

the first floor of the early MCC but displayed to operators in the new facility. This last residual function will be phased to 30S by 2002.)

The latest step in contracting for MCC services also occurred in 1998, when NASA awarded the Consolidated Space Operations Contract (CSOC) to Lockheed, which had acquired Loral, which had earlier acquired Ford Aerospace. To complete the contracting history (see Fig. 12), the Rockwell contract for STSOC had been reassigned to a joint venture of Rockwell (later Boeing) and Lockheed known as United Space Alliance (USA) before the MCC sustaining and operating segments of that contract were transferred to the new CSOC.

Perspective

For the past 40 years, the operating concept of a ground-based flight control team has been very effective. The concept envisioned the team responding to program objectives and staffed with all of the necessary disciplines for real-time decision making. The concept has been driven by the need for rapid and critical response to flight situations, within seconds when necessary. This has led to careful selection of personnel and a widely scoped training regimen from classrooms to fully integrated simulations with the flight crews in simulators. An early emphasis was established for flight controllers to develop as many of their own tools as possible, from functional schematic systems handbooks, to mission rules and procedures, and to the requirements for all of the processing, displays, and controls on their respective consoles. In-depth support to each console operator from staff support rooms has also been a constant feature. With variations for program content and mission phases, the concept and the makeup of the flight control team has been remarkably consistent for four decades.

In terms of interfaces with other operational units, the flight control team defined the technical requirements (voice, data, and other) for any transactions with these remote units and the procedures governing same. This has led to the

flight control team in the MCC acting as the central hub on the ground for mission support with interface connections to units such as the launch team, range safety, network and MCC systems, DOD support, and payload organizations in their control centers at GSFC, JPL, MSFC, the Air Force Sunnyvale center, or in other facilities. A very strong interface with the spacecraft program office team of NASA/JSC engineering and industry teams has also evolved. This interface assured maximum confidence in resolving technical problems and a controlled methodology for engaging the program office team. Because the real-time flight control team has been augmented by and connected to all the necessary organizations, it has been a vital element in all of the achievements and successes of human spaceflight.

The MCC is the instrument by which this operating concept has been mechanized. Its fundamental architecture has three elements—front-end communication processing, a computation complex (now servers), and a display and control element (now work stations with local processing added). It has evolved over the years in equipment and configurations but has provided the same fundamental core functions.

After nearly perfect support to 114 missions by the original MCC, its third floor FCR is now maintained as a national historical monument commemorating Apollo 11. The second floor has been modified to support life science payloads. The original Building 30 MCC had performed nearly flawlessly as NASA achieved its human space goals during the last four decades. It is now retired with honor from its original mission, and the new MCC is poised to write its own chapter in human space history.

BIBLIOGRAPHY

1. Legler, R.D. (For various historical numbers included in text) *Responses to Questions About Historical Mission Control April 97*.
2. Loree, R. *MCC Development History*, August 1990.
3. Stokes, J.C. *Managing the Development of Large Software Systems*. Apollo Real-time Control Center, August 1970.

GLYNN S. LUNNEY
Houston, Texas

NUCLEAR ROCKETS AND RAMJETS

Introduction

The objective of this article is to describe the design concepts of and the programs to develop our nuclear rocket propulsion capability for spaceflight and the nuclear ramjet systems intended for missile delivery and to describe the mission capabilities they could provide. In both systems, the nuclear reactor replaces the

combustion chamber as the source of high-temperature energy used to provide the thrust in place of their chemical alternatives. The nuclear rocket was, and still is, considered for deep space missions, including missions, such as human flight to Mars, whereas the stimulus for the nuclear ramjet was its potential performance as a high-speed, almost limitless range, low-altitude and, therefore, almost undetectable missile delivery system. In the early phases of research on the nuclear rocket, it too had been considered as a possible missile system.

A drawing of the nuclear rocket propulsion system is shown in Fig. 1, including identification of the major components of the system. In this system, low molecular weight hydrogen is heated to high temperatures in the nuclear fission reactor to produce a specific impulse that is about twice (900 seconds compared to the 452 seconds of the hydrogen-fueled combustion Space Shuttle Main Engine) and could eventually be even much greater than the level of the best chemical rocket systems. This system also required the advanced development of large capacity hydrogen turbopump systems and also hydrogen-cooled jet nozzles that operate at the high temperatures of the jet exhaust hydrogen gas; this technology was not yet available in the early phases of the program. In the nuclear ramjet, which is shown with its major components in Fig. 2, the air flowing through the system, as a result of velocity imparted to the vehicle by a separate launch system, is also heated to high temperature in the nuclear reactor. No propellant fuel must be supplied and carried, so it could have a very long range. However, the inlet system must be designed to reduce pressure losses efficiently through the inlet shock waves and to reduce the air velocity through its inlet diffuser from its supersonic entry to subsonic levels so that the air can be heated in the reactor core and then accelerated efficiently in the jet nozzle. Those systems and, of course, high-temperature reactors were not yet proven when the program started. However, analysis and technology development was underway in various laboratories on the reactors, pumps, diffusers, and nozzles for rocket and ramjet systems.

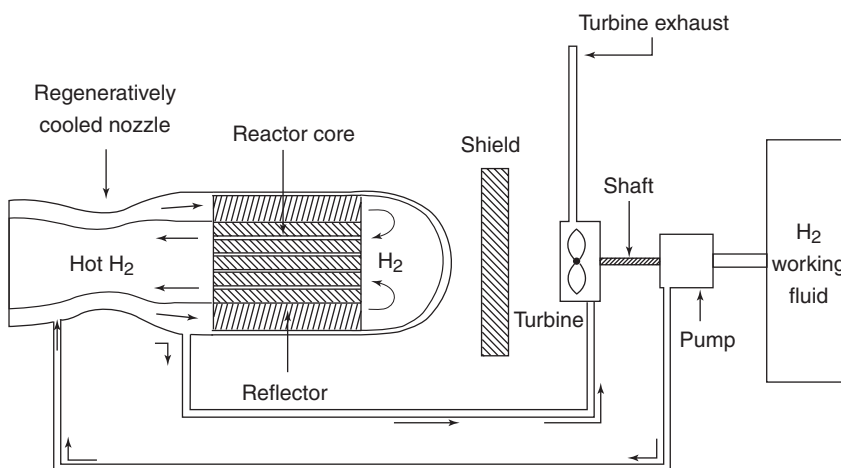


Figure 1. Nuclear rocket engine.

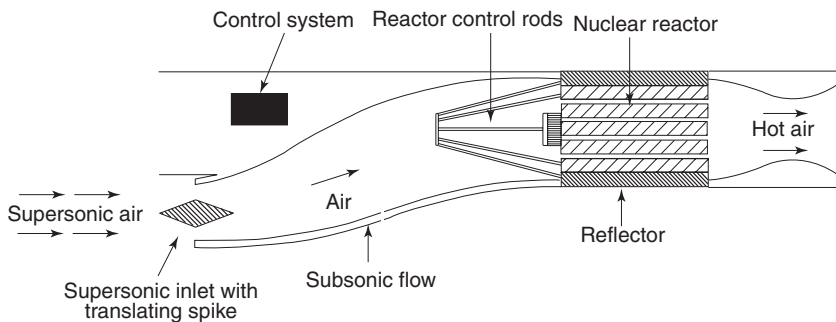


Figure 2. Nuclear ramjet propulsion system.

Of course, in both systems, the hardware weight of the entire system must be limited so that the various mission benefits resulting from the nuclear energy source are not counterbalanced by that total weight. Major design and development requirements include selecting materials that have high-temperature strength and are compatible with hydrogen propellant in the rocket and air in the ramjet, appropriate neutronic characteristics, controlled and stable start-up and operation, component design and reliability, and are safe. All of these matters must be and were comprehensively researched and developed to provide systems that could achieve the desired performance. These were all high priority tasks in both the nuclear rocket and nuclear ramjet programs, and they are discussed in this section, as are the important management organizations and technical capabilities established to carry them out.

Though significant progress was made in these programs, neither the nuclear rocket nor the nuclear ramjet program was carried to actual application in its proposed and anticipated missions. In the nuclear rocket case, the space program has not yet committed to mission objectives that require the high payload and deep space capability of that system. However, nuclear rocket propulsion continues to be recognized as necessary for large, deep space missions. In the nuclear ramjet case, existing chemical rocket systems provide all of the military missile requirements that are foreseen. Similar to the nuclear ramjet development, the extensive work that had been done on nuclear-powered turbojet engines, that is mentioned in this section, was also discontinued because chemically fueled turbojets provide all of the capability that was required.

Origins of The Nuclear Rocket and Ramjet Programs

The twentieth century produced an outstanding array of scientific and technological developments that have had major effects on our planet, on the economies of the various nations, on national security, and on individual lives around the world. Several of them combined to provide the basis for establishing the theoretical and experimental background and the broad capabilities required to achieve the developments and missions discussed throughout this encyclopedia and, very specifically, the flight propulsion developments in this section. Among

these scientific and technological innovations that were the originating stimulus and foundation for the concepts of nuclear rocket and ramjet propulsion were Orville and Wilber Wright's first powered aircraft flights on 17 December 1903 at Kitty Hawk, North Carolina (1); Robert Goddard's concepts and development of his first rocket flight on 16 March 1926 and his continuing developments and flights beyond that (2), and then, of course, the discovery of the nuclear fission process in the late 1930s and the demonstration of the first controlled nuclear chain reaction on 2 December 1944 under the grandstands of Stagg Field Stadium at the University of Chicago that "marked the birth of the nuclear age" (3,4). Certainly, these achievements were the stimulus for applying nuclear energy to flight systems.

But it must also be recognized that many of the basic concepts that led to these developments were originally theorized and suggested and, in some cases, even purely imagined by diverse individuals and organizations throughout the United States and in many other nations throughout the world. The continuing research and development of these and many other individuals and organizations in and from various nations generally led to achieve the spaceflight capabilities described here. For example, as P.E. Cleator (5) points out in his 1936 book entitled *Rockets Through Space – The Dawn of Interplanetary Travel*, some were even imaginatively fictional space propulsion conceptions that encouraged suggestions of advanced concepts to overcome Earth's gravity—such as Jules Verne's shot into space by a huge cannon and H.G. Wells' "conveniently discovered" substance he called Cavorite "to which the earth was definitely repellant." Both of these citations also indicate the long fascination with travel in space. P.E. Cleator suggested some of these very advanced and imagined space propulsion concepts, including improved fuels and the use of solar energy and, even before Stagg Field's chain reaction, "atomic energy" recognizing "that matter contains tremendous stores of potential energy, if only we knew how to release and use it." He used the 10-fold increase in energy released by the disintegration of radium per gram compared to that of coal as the basis for his "fanciful and remote—possibility of utilizing atomic energy." In 1943, Robert H. Goddard speculated on the use of "inter-atomic energy" so a "large body could be sent from the solar system—after the problem of atomic disintegration has been solved" (6). Many others among the early scientists and those who were involved in early nuclear physics research perceived realistic possibilities of atomic energy.

The fundamental discoveries and technological achievements within the first half of the twentieth century stimulated advancements that led to the development of turbojet aircraft engines, to extensive work on nuclear aircraft propulsion, to chemical rocket and airbreathing missile development, and to various nuclear-powered propulsion concepts, including the nuclear rocket and ramjet systems discussed here. And, of course, from the objective of achieving deep spaceflight, a host of other innovative propulsion energy concepts were also identified, including ion propulsion, plasma propulsion, gas core and particle bed nuclear rocket reactors, nuclear bomb propulsion (the Orion Project), solar sails, and antimatter propulsion; some of them are still being explored today.

It must also be recognized that in the 1940s and 1950s, we were at war. In its aftermath, various international antagonisms, crises, and strong competition emphasized and stimulated the need to demonstrate visibly our international

technological leadership to ensure national security. Certainly, the Soviet Union's launch of its Sputnik I in October, 1957 contributed greatly to this need. President Eisenhower referred to that event when he said, "whatever our hopes may be for the future—for reducing this threat or living with it—there is no escaping either the gravity or the totality of its challenge to our security and survival—a challenge that confronts us in unaccustomed ways in every sphere of human activity." Then on 25 May 1961, only a month after Yuri Gagarin's first manned orbital flight, President John F. Kennedy delivered his famous "Special Message—on Urgent National Needs" (7) to a joint session of the Congress. In that message, he proposed "that this nation should commit itself to achieving the goal, before this decade is out, of landing a man on the moon and returning him safely to the earth." His second space proposal was to "accelerate development of the Rover nuclear rocket" which "gives promise of some day providing a means for even more exciting and ambitious exploration of space, perhaps beyond the moon, perhaps to the very end of the solar system itself."

Existing Capabilities in the 1950s

Achieving and ensuring such internationally visible technological strength and preeminence required the active involvement of many organizations that could provide sound mission analysis and research and technology development in areas of basic technical and scientific knowledge, such as those mentioned before, to achieve the planned goals. Fortunately, many of these capabilities already existed in the 1950s. They included leadership, support, and technical and management capabilities of the various military organizations, including the U.S. Army, the Navy, and the Air Force (USAF) that depended on aircraft systems and missiles to fulfill its national security responsibilities; the National Advisory Committee for Aeronautics (NACA) which the 1915 Naval Appropriations Bill had created to conduct research on aircraft systems for "the scientific study of the problems of flight, with a view to their practical solution" (8); NACA became the foundation of the National Aeronautics and Space Administration (NASA) when that organization was established on 1 October 1958 in the U.S. response to the Soviet Union's demonstration of its space capability, and the U.S. Atomic Energy Commission (AEC) which was created under the Atomic Energy Act of 1946 (9) to provide civilian control over the development of nuclear energy, including military as well as peaceful civilian applications. President Truman's Executive Order transferred control of the nuclear weapons development program from the Army to the AEC on 1 January 1947. The AEC later became part of the Energy Research and Development Administration and then the Department of Energy.

Progress in developing nuclear rocket and ramjet propulsion was the result of research, design analysis, and technology development led and carried out by technical organizations within these various government agencies, generally working collaboratively among themselves but also working very closely with private companies and with universities. The government laboratories included some of those that had previously been established to develop our first atomic bombs, such as the AEC's Los Alamos Scientific Laboratory, the Lawrence Livermore Radiation Laboratory, the Oak Ridge National Laboratory, and the

Argonne National Laboratory. In addition, the NACA/NASA laboratories, such as the Langley and Ames Research Centers, had been heavily involved in high-speed aircraft research; the Dryden Flight Research Center had been doing high-speed aircraft flight testing; and the Lewis Research Center (recently renamed the John H. Glenn Research Center) had been working on propulsion systems, including turbojet, rocket, and ramjet engine research and the hydrogen components that would go with advanced rocket engines, well before the space program was formally initiated (10). That Laboratory had originally been named the NACA Aircraft Engine Research Laboratory when George W. Lewis broke ground for it on 23 January 1941 (8). The Wright-Patterson Air Force Base was involved early in military aircraft development. The work of these various research and development organizations was generally conducted in close collaboration with the development and engineering capabilities of many companies throughout the nation. Among these were the Aerojet General Corporation; Westinghouse Electric Corporation; Rocketdyne Division and Atomics International of North American-Rockwell Corporation; General Electric Company; Glenn L. Martin Company which later became part of Lockheed Martin; Pratt & Whitney Aircraft which later became part of the United Technology Corporation; Thiokol; the DuPont Company; Edgerton, Germeshausen, and Greer; General Atomics; Marquardt; Chance-Vought and many other companies and government and university laboratories. The collaborative efforts and capabilities of these various organizations—government, industry, and universities—have made advanced deep space missions realistically achievable when decisions are made to commit to them. Several of those companies also continued to play a major role in developing and providing our U.S. aircraft and missile system capability.

Management Arrangements for Nuclear Flight System Programs

As a result of the interests of the USAF and its discussions with Dr. Vannevar Bush of the Office of Scientific Research and Development, as well as with Major General Leslie Groves, who headed the Manhattan Project's atomic bomb development, studies of the feasibility of nuclear-powered propulsion for aircraft were undertaken (11). These led the Pentagon to establish, in 1950, an Aircraft Nuclear Propulsion Program (ANP), which was conducted as a joint effort of the AEC and the Air Force. About a year later, an Aircraft Nuclear Propulsion Office (ANPO) was organized within the AEC staffed jointly by the AEC and the Army Air Corps with Major General Donald J. Keirn as its Director. Brigadier General Irving L. Branch succeeded General Keirn in 1959. Consistent with the Atomic Energy Act provisions and intent, the AEC provided the nuclear phases of the work and the USAF was responsible for the nonnuclear and aircraft aspects, but all were managed by this new, single, joint organization. At that time, the objective of the program was raised from evaluating the feasibility of nuclear-powered aircraft flight to include actual demonstration of nuclear-powered flight. General Electric and Pratt & Whitney (which was working with the AEC's Oak Ridge National Laboratory) were under contract to conduct research and development on direct- and on indirect-cycle concepts of nuclear aircraft propulsion, respectively, for the ANPO.

During that period, in the first half of the 1950s, Los Alamos and Livermore were also involved in conducting work on nuclear rocket propulsion using non-project research funds available to them (12, pp. 3–6). In response to an inquiry from the AEC in September 1955, the Department of Defense (DOD) “indicated an intense interest in this program” but on the basis that it not interfere with the Los Alamos and Livermore nuclear weapons programs. In essence, that was interpreted by the ANPO to mean proceeding with a ground test and development program “to determine the feasibility by ground operation of a nuclear rocket engine by about 1959, looking forward to a possible flight in (deleted) if this appeared desirable.” A formally designated nuclear rocket propulsion program was then established under the management of the ANPO, and budget funding was requested. However, in 1956, Secretary Wilson of the DOD sent a letter to Chairman Strauss of the AEC modifying that requirement and removing the sense of urgency that had been previously indicated by asking the AEC to “continue on a moderate scale to develop a reactor suitable for nuclear propulsion of missiles, satellites, and the like” (12).

As a result, “funds were reduced and a decision was made to go to a single laboratory approach” for the nuclear rocket propulsion program. Los Alamos continued its work on the nuclear rocket program, whereas Livermore dropped its nuclear rocket work and, in January 1957 (12), was assigned work on the nuclear ramjet program called Project Pluto. Both of these programs are the major subjects of this article. The Rover name that Livermore had used for its nuclear rocket work replaced the Los Alamos Condor designation for the nuclear rocket program. The ANPO provided the overall management of both the nuclear rocket and the ramjet projects and, also, had responsibility for developing radioisotope and reactor space power systems under its Missile Projects Branch.

While these actions were being taken by the AEC, the Navy, Army, and Air Corps were conducting major activities to develop advanced missiles and research on using advanced fuels (13). At the same time, during the 1940s and 1950s, the NACA laboratories were continuing their research, mentioned before, to achieve high-speed aircraft flight, including vehicle aerodynamic advancement and advanced propulsion systems work that involved work on turbojet, rocket, and ramjet systems at the Lewis Research Laboratory. That work was regularly reported and presented to broad government and industry audiences (10) in classified (since then declassified) conferences during the 1940s and 1950s, well before NASA was established, and conferences continued after NASA was formed.

When NASA was formally activated on 1 October 1958, responsibility for the nuclear rocket’s nonnuclear components, subsystems, and integrating them with the AEC nuclear reactor into the engine system and in flight vehicle development was transferred by Executive Order from the USAF to NASA (11). NASA staff was assigned to work with the ANPO, Los Alamos, and other participants in the program to fulfill NASA’s responsibility. Recognizing the AEC’s legislatively assigned responsibility for nuclear reactor systems research and development, it was proposed soon after NASA was established, that a joint office of the AEC and NASA should be formed that would carry out the Rover Program’s reactor and nonnuclear component development and their application to engine systems. However, as is indicated in his diary (14) “The Birth of NASA,” it

took Dr. T. Keith Glennan's—NASA's first Administrator—persistent effort and great patience to establish the joint office. In all such AEC organization as well as program and budget arrangements, the Joint Committee on Atomic Energy (JCAE), which was established by the Congress in 1947 to provide congressional oversight of all AEC activities, was very actively involved and played powerful and influential roles. Dr. Glennan had experience with the AEC and JCAE because he had served as one of the five Commissioners from 1950 to 1952 (15), and in 1955, he began serving on a committee under the JCAE direction (14).

The major obstacle in establishing the joint NASA and AEC office related to selecting the person to head that new joint office that would replace the ANPO as manager of the nuclear rocket program. Members of the JCAE including the powerful Senator Clinton P. Anderson (14) and some others (16) favored the appointment of Air Force Colonel Jack L. Armstrong. He was then the Deputy Chief of the ANPO and was well known to the JCAE. NASA favored Harold B. Finger, the author of this paper, who had been a research engineer at the NACA Lewis Research Center starting in May 1944, and was working in the NASA Headquarters on nuclear systems; he was the key person who had been assigned as the link with the AEC–Air Force ANPO on the nuclear rocket program when NASA was established.

Finally, Dr. Glennan and the AEC Chairman John A. McCone signed a space nuclear rocket Memorandum of Understanding on 26 August 1960 and 29 August, respectively (17), and on 31 August 1960, the joint office was publicly announced by the two agencies. The author was named Manager of the AEC–NASA Nuclear Propulsion Office, and Mr. Milton Klein, who had been Assistant Manager for Technical Operations of the AEC's Chicago Operations Office, was designated as his Deputy. In that Memorandum of Understanding and in the public announcement, the responsibilities of the two agencies were clearly defined. The AEC had primary responsibility "for conducting research and development on all types of nuclear reactors and reactor components, including those required for aeronautical and space missions specified by NASA," and NASA had "primary responsibility for research and development on nonnuclear components and integration of the nuclear components in engines and vehicles of rocket systems." At that point, the nuclear rocket responsibility was transferred from the ANPO to the new AEC–NASA joint office, and key people, including two USAF officers, were transferred to the new office. Therefore, experienced people from the Air Force, the AEC, and NASA staffed the office.

That joint office was very soon more descriptively renamed the AEC–NASA Space Nuclear Propulsion Office (SNPO). That clearly emphasized the objective of its work and distinguished it from the original consideration of potential ICBM applications that had stimulated the Air Force's original interest in nuclear rocket propulsion but was canceled by the Defense Department in 1956. The resulting AEC and NASA organizational structure is shown in Fig. 3. The Nevada, Albuquerque, and Cleveland Extensions shown on the chart were established by the SNPO to provide its management support for on-site activities and, in Cleveland, for managing its engine contract work. The Cleveland Extension drew on the technical support of the Lewis Lab because of its experience and continuing work on hydrogen pumps, nozzles, and other components needed for

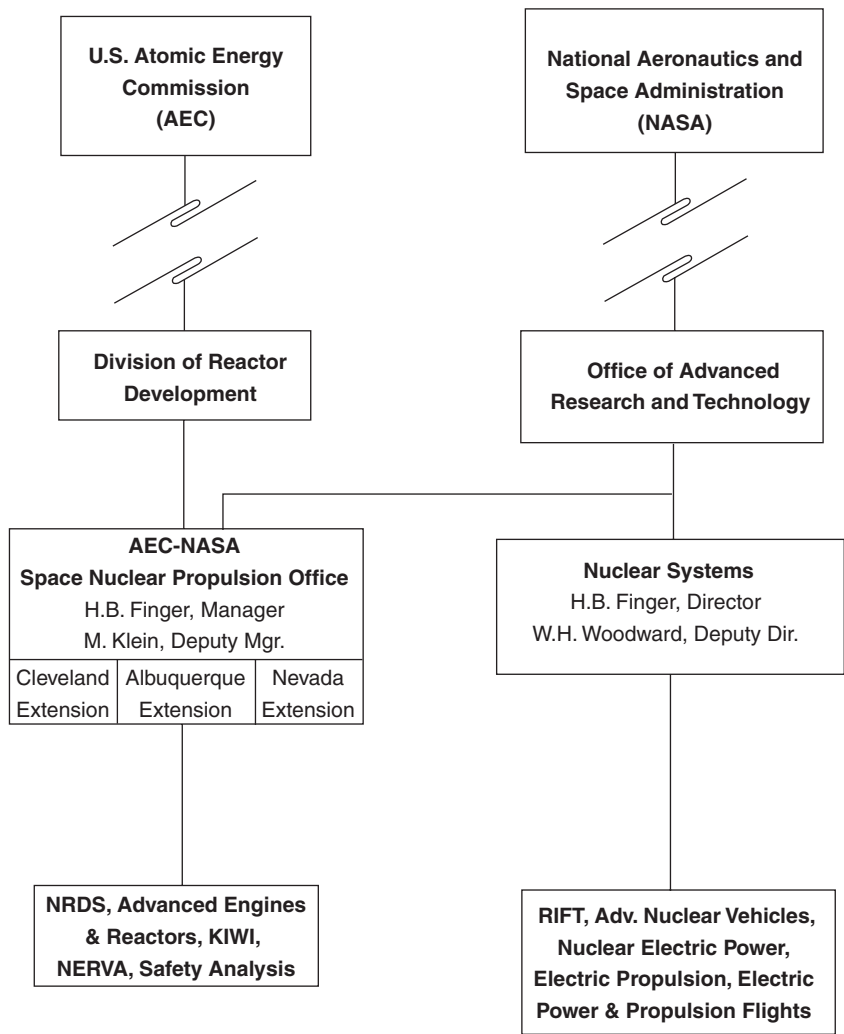


Figure 3. Organization—nuclear activities.

the nuclear rocket engine as well as reactor support. In 1965, the author was also named the AEC’s Director, Space Nuclear Systems Division to head the AEC’s space nuclear power systems development.

Mission Performance Benefits of Nuclear Rockets

Although the nuclear rocket program was started by the Air Force as a possible approach to strengthening its potential missile capabilities, it was also considered very early by those identified before as a possible space launch system. Important mission benefits were and still are anticipated from nuclear rocket propulsion systems compared to chemical rockets, if nuclear systems could be

satisfactorily developed. For example, the statement of Dr. Hugh L. Dryden, Deputy Administrator of NASA (18), on Project Rover at the 27 February 1961 Hearings of the House of Representatives Committee on Science and Astronautics emphasized this point. "We in the NASA are particularly aware of the large advantages that nuclear energy offers in our space program.—Our evaluations have made it clear that the space program will require the application of nuclear energy sources in order to provide the large amounts of energy that are required to move about freely in space.—We are confident that—probably the first means will be chemical, but when you get to transporting large amounts of material, nuclear energy is essential—manned exploration of the moon and the planets rests on the mission capabilities afforded by nuclear propulsion systems. For this reason, we consider the development of nuclear propulsion systems as our major advanced propulsion development program." And, on 28 August 1961, shortly after he was appointed NASA's second Administrator, Mr. James E. Webb stated in testimony before the Joint Committee on Atomic Energy (19), "We look to the nuclear rocket primarily for application to missions beyond the first manned lunar expeditions; for providing the heavy payloads that may one day be required to support lunar bases and for manned exploration of the planets. Nuclear energy is essential for such missions." That is still the case.

Even before Wernher von Braun and his rocket team in the Army Ballistic Missile Agency were being phased into NASA, starting in November 1959 and culminating in March 1960 when they became the NASA Marshall Space Flight Center, they had already been deeply involved in developing the large Saturn launch vehicle which various NASA analyses (20) had indicated was necessary for its eventual space missions, whereas the Department of Defense Advanced Research Projects Agency (ARPA) could not anticipate missions that would require such a large launch vehicle. During that time, the Manager of the SNPO emphasized to members of the von Braun team and in meetings, including the Administrator, Deputy Administrator, and others in NASA Headquarters, that the Saturn vehicle should be made as large in diameter as possible in order to ensure that it could accommodate nuclear-powered upper stages containing the large volume hydrogen tanks that would be required. As a result, the diameter was set as large as was permitted by the hook height in the vehicle assembly building at the Marshall Center. NASA emphasized (19) that its chemical rockets were being designed so that nuclear upper stages could be applied to them to increase "payloads for various missions including lunar missions."

Figure 4 (18) presents the increase in escape payload for one of the early Saturn vehicle concepts by using a nuclear third stage on the first two stages of the Saturn C-2 configuration, compared to a three- and a four-stage chemical Saturn C-2. (The Saturn C-2 was one of several configurations being examined at the time this figure was originally presented.) The upper level of the shaded Saturn line is for the four-stage chemical vehicle, and the lower level is for the three-stage vehicle. Even at a reactor thermal power of 1000 megawatts, which would provide about 50,000 pounds of thrust, the nuclear system would deliver an escape payload of twice the all-chemical system. As the reactor power and resulting thrust increase, that increase in payload grows significantly so that at 4000 Mw, the escape payload of the nuclear system is three times that of the all-chemical Saturn.

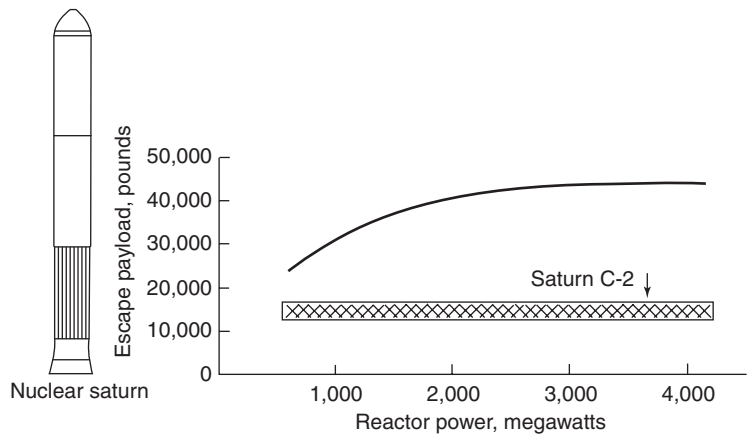


Figure 4. Nuclear stage increases Saturn escape payload.

Figure 5 compares the payload that could be delivered in various unmanned solar system missions with a chemical or a nuclear third stage on the first two stages of the large Saturn V vehicle which served as the first stages of the launch system for the Apollo mission. The nuclear third stage vehicle provides at least twice the all-chemical system payload. The benefits of nuclear propulsion are far greater for human missions. To accomplish a human round trip mission to Mars, the total weight required in Earth orbit with a nuclear powered vehicle is less than half that of the best chemical rocket in the most favorable Mars–Earth planetary alignment with a significant reduction in trip time. That advantage of nuclear propulsion increases substantially at other less opportune times since the Earth orbital weight required for the chemically propelled vehicle’s Mars

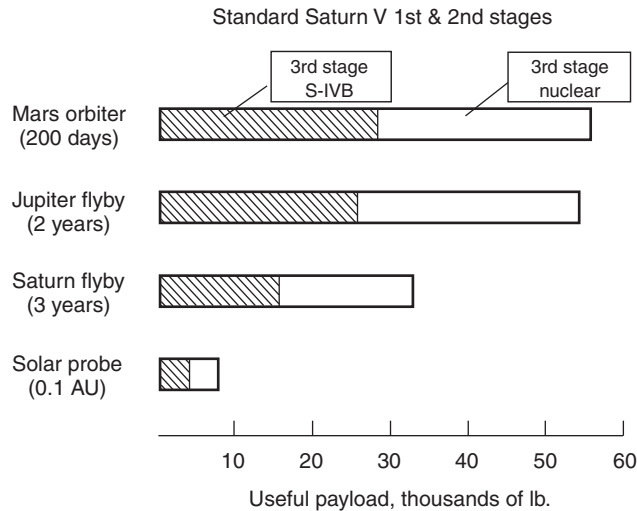


Figure 5. Unmanned solar system missions—application of nuclear third stage to Saturn V.

mission accomplishment increases significantly and rapidly compared to the nuclear propelled mission.

These performance benefits can be summed up as they were reemphasized in the 1991 report (21) of a Synthesis Group on America's Space Exploration Initiative chaired by Thomas P. Stafford, which pointed out the important role of nuclear thermal rocket propulsion in space exploration and singled out Mars exploration:

Nuclear thermal propulsion has approximately twice the performance of chemical rockets, with reduced propellant requirements. This leads to reduced mass in low Earth orbit, faster trip times (increasing crew safety) and increased launch windows.

The Nuclear Rocket Rover Program

Formally designated on 30 August 1961, the SNPO initiated major program actions, and the Los Alamos Scientific Laboratory (LASL) continued its responsibility and effort for the nuclear rocket reactor program aimed at developing and testing its nuclear reactors in the facilities that had been built under Los Alamos direction at our Nuclear Rocket Development Station (NRDS) in the Jackass Flats area of the AEC's Nuclear Test Station in Nevada. Those early reactors were named after the flightless KIWI bird. When the SNPO was established, Los Alamos had already tested two reactors, the KIWI-A on 1 July 1959 and the KIWI-A' on 8 July 1960 (both discussed later).

The early SNPO and NASA program actions are listed in Fig. 6. They indicate the broad scope of the program and the goals of full nuclear rocket engine and flight vehicle development aimed at flight testing. RIFT was the Reactor In-Flight Test rocket stage; NERVA was the Nuclear Engine for Rocket Vehicle Application; the MAD Building was the engine Maintenance and

1. 31 August 1960: AEC-NASA Space Nuclear Propulsion Office established.
2. September 1960: Contracts with Convair, Douglas, Lockheed, Martin for flight testing nuclear rockets (RIFT).
3. December 1960: Contract with Parsons team on master plan for required nuclear rocket engine development facilities.
4. February 1961: Issued RFP for NERVA contractor. Proposals due April 3.
5. 7 June 1961: Aerojet General-Westinghouse Team selected for NERVA contract.
10 July 1961: NERVA contract signed.
6. July 1961: RIFT studies extended.
7. August 1961: Contract with Vitro to design Engine MAD Building. Construction started in 1962.
8. 11 July 1962: RIFT development contract awarded to Lockheed.

Figure 6. Early Space Nuclear Propulsion Office Program actions.

Disassembly Building that was required, so that the system disassembly and any modifications after ground testing could be done using shielded remote manipulator operation. The AEC and Los Alamos had already constructed a reactor MAD building, and it had already been used to disassemble the KIWI-A and A' reactors after Los Alamos had tested them.

The Reactor and Engine Development Test Program. The sequence of the resulting reactor and engine development testing activities is shown in Fig. 7, and the operating time and power level are given in Fig. 8. All of these tests were conducted at the NRDS in Jackass Flats which included two reactor test facilities, an engine test stand equipped to permit test operation under space vacuum conditions, and a reactor and an engine maintenance and disassembly building.

As pointed out in the Introduction of this section of the Encyclopedia, a major issue involved in designing the nuclear rocket reactor was and is the choice of materials to withstand the high temperatures required and to ensure compatibility with the hydrogen propellant and among the various other materials in the system. In addition, the neutron capture cross-section of the materials becomes a significant property in determining the type of reactor and fissionable fuel loading required. Numerous analyses and tests have been conducted on various potentially suitable materials for rocket reactor design by various participants involved in research and analysis of nuclear rocket propulsion. Among them are the work of R.W. Bussard (22) of LASL, Frank E. Rom (23) of the Lewis Research Center, and Donald P. MacMillan (24) of LASL. Among the materials considered have been graphite as both the fissioning fuel element and neutron-moderating material, tungsten, molybdenum and others as well as various carbides for the fuel material with beryllium or beryllium oxide for moderator and reflector materials. The Argonne National Laboratory became involved in research on tungsten-fueled fast neutron reactor concepts, and the Lewis Research Center worked with reactor systems using tungsten fuel elements and water moderation systems that required the enriched tungsten-184 isotope to reduce the high neutron capture cross-section of natural tungsten (23).

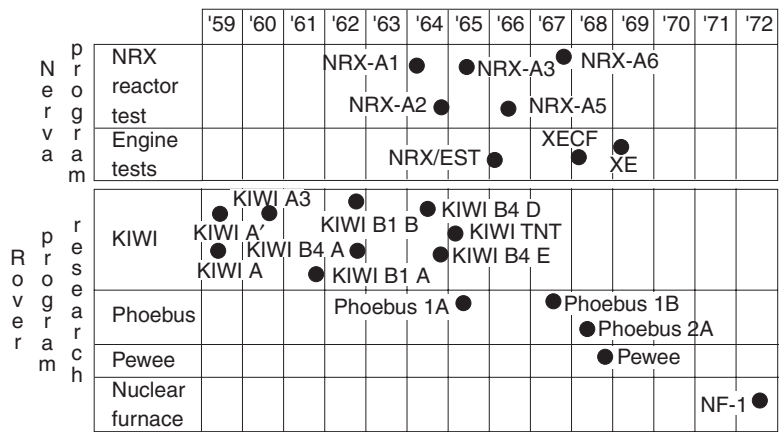


Figure 7. NERVA/Rover reactor system test sequence.

Date	Test Article	NRDS Test Facility	Maximum Power	Time at Maximum Power
1 July 1959	KIWI-A	A	70 MW	5 min
8 July 1960	KIWI-A1	A	85 MW	6 min
10 October 1960	KIWI-A3	A	100 MW	5 min
7 December 1961	KIWI-B1A	A	300 MW	30 sec
1 September 1962	KIWI-B1B	A	900 MW	Several sec
30 November 1962	KIWI-B4B	A	500 MW	Several sec
13 May 1964	KIWI-B4D	C	1000 MW	40 sec
28 July 1964	KIWI-B4E	C	900 MW	8 min
10 September 1964	KIWI-B4E	C	900 MW	2.5 min – restart
24 September 1964	NRX-A2	A	1100 MW	40 sec
15 October 1964	NRX-A2	A	Restart	(mapping)
23 April 1965	NRX-A3	A	1165 MW	3.5 min
20 May 1965	NRX-A3	A	1122 MW	13 min
28 May 1965	NRX-A3	A	≤500 MW	1.5 min (–28.5 min)
25 June 1965	Phoebus 1A	C	1090 MW	10.5 min
3, 16, 23 March 1966	NRX-EST	A	1100 MW	1.5 min – 14.5 min – 8 min
23 June 1966	NRX-A5	A	1140 MW	1.5 min – restart – 14.5 min
23 February 1967	Phoebus 1B	C	1500 MW	30 min
13 December 1967	NRX-A6	C	1100 MW	62 min
26 June 1968	Phoebus 2A	C	4200 MW	12 min
3–4 December 1968	Pewee	C	514 MW	40 min
11 June 1969	XE-Prime	ETF-1	1100 MW	11 min
29 June–27 July 1972	Nuclear Furnace	C	44 MW	109 min (4 tests)

Figure 8. Chronology of Rover and NERVA reactor/engine tests.

After considering the various high-temperature materials alternatives in its extensive early research and development work that started in 1955, LASL selected graphite in spite of the fact that, unless effective protective coatings were found, the hydrogen reactor coolant and rocket propellant would severely corrode the graphite and form methane or acetylene at the high operating temperatures required for successful nuclear rocket operation. Materials research including such coatings and protective concepts were an important part of the KIWI and follow-on Rover research and development program. The graphite reactor systems were the principal reactor development objective throughout the Rover program, including both the Los Alamos KIWI reactors and the Westinghouse reactors for the NERVA rocket engine, which was based heavily on the Los Alamos KIWI work, and for the later, higher power LASL Phoebus reactor designs. Although work continued on other reactor concepts, including the tungsten fuel element systems mentioned before and also gas core reactor work, the major emphasis through the entire program was on the graphite reactor development effort conducted by Los Alamos and then the graphite reactor-engine system work by Aerojet-General and Westinghouse. The ultimate outcome of that work was successful achievement of all of the objectives that had been set. That put this country in a solid position to undertake flight system development and mission accomplishment with a high degree of confidence when required missions are defined. However, such missions have not yet been defined.

The KIWI-A Reactors. All of the KIWI-A series reactor tests (A, A', and A3) listed in Figs. 7 and 8 were aimed at about 100 MW thermal power levels. Work directed toward those test reactors had started in the mid-1950s. All of those

tests were run with gaseous hydrogen stored in high-pressure tanks as the reactor coolant and used water-cooled Rocketdyne nozzles. Their purpose was primarily to check out basic elements of the reactor operations and design and to serve as facility checks. Though problems were encountered, including some that were anticipated, the Kiwi-A series demonstrated that high hydrogen temperatures could be produced in these heat transfer reactors with controlled start-up, stable operations, and shutdown.

Dr. Raemer E. Schreiber, Chief of N-Division in Los Alamos and responsible for the Rover Program, described the first test of the KIWI-A (12), as “a test device for our own education in order to get us the first information on an integral system which has some of the characteristics which we are looking for in actual propulsion engines.—The relationship between this and a flyable device is pretty tenuous.” That KIWI-A reactor test system is shown in Fig. 9 on the railroad car moving it to Test Cell A from the reactor assembly building. All of the reactor tests were fired upward. The KIWI-A reactor system (25) consisted of uranium-235 (UO_2)-loaded graphite fuel plates in an annular section around a central section containing heavy water (D_2O) and the rods that controlled the fission process, including those that could scram the reactor in an emergency. The fuel element plates were 1/4 inch thick and were spaced 0.050 inches apart to provide for the hydrogen flow. This first test had no protective coating on the graphite fuel. Although part of the center island blew out at start-up, the test was run for about 5 minutes, limited by pressurized hydrogen capacity, up to a power level of 70 MW. Because no corrosion protection was used, it was not surprising that heavy corrosion occurred, but those results “showed no disagreement with laboratory results” (25). However, in addition, most elements were cracked and part of the central island blew out through the nozzle during start-up.

The configurations of the second and third test reactor, the KIWI-A' and the KIWI-A3, were very similar to each other but very different from KIWI-A. The

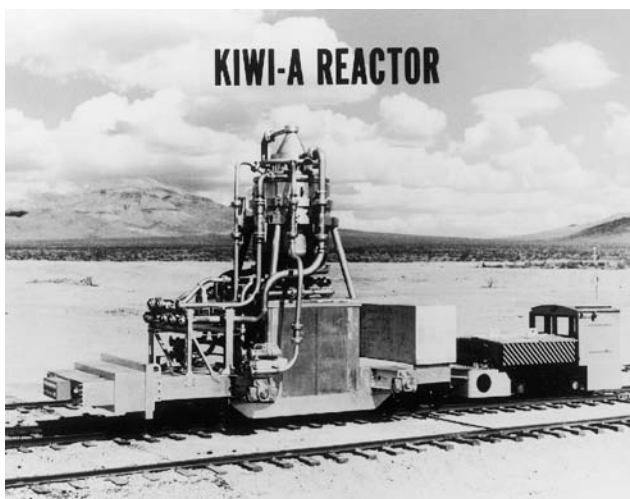


Figure 9. KIWI-A Reactor in transit to Test Cell A.

KIWI-A' test was run on 8 July 1960 using a cylindrical fuel element configuration that began to approach the hexagonal fuel elements planned for future reactors. However, the KIWI-A' fuel elements had four flow passages through each fuel element rather than the 19-hole hexagonal fuel elements in the later KIWI-B liquid hydrogen reactor test systems. The KIWI-A3, which was run on 19 October 1960, had seven flow passages (26) in each of its long cylindrical fuel elements. Both of these reactors had hydrogen corrosion protection applied through chemical vapor deposition of niobium carbide (NbC) on the fuel element flow passages. The KIWI-A' test ran for about 7 minutes up to a power level of 85 MW. Seven of its fuel elements broke, and all others showed heavy cracking. The protective coating looked satisfactory except for some blistering. The KIWI-A3 test ran for about 5 minutes up to 100 MW. Remote disassembly and examination after the test indicated less corrosion than KIWI-A': one element broken and significant fuel cracking. Though these various KIWI reactor tests indicated significant inadequacies, they also provided design, materials, structural, control, and operating data that were useful in the continuing development of the higher power reactors that were to follow.

The KIWI-B Reactors. Although the KIWI-A series was aimed at about 100 MW thermal power, the KIWI-B series (there were several designs being examined at that time) was designed for 1000 MW (27). The goal was achieving as high an operating temperature as possible through the development program and a high power density. The KIWI-B reactors were to serve as the basis for the NERVA engine reactors (25) to provide an engine thrust of 50,000 pounds and the high specific impulse expected of nuclear rocket propulsion. Although much was learned from the KIWI-A series, the KIWI-B reactors were significantly different. The KIWI-B' used a beryllium reflector, in which the control rods were located, around the periphery of the reactor core. After detailed examination of the different core designs, including the structural results of the KIWI-A test series and considering facility availability, it was decided first to test the KIWI-B1 with gaseous hydrogen at reduced power of 300 MW. The fuel elements in that reactor were similar to the seven flow-hole cylinders in KIWI-A3 and were located in a graphite matrix (26). The operating time and power of that test (Fig. 8) were limited by an emergency reactor scram and a fire caused by a leak in the seal between the nozzle and pressure vessel. In spite of serious concerns about the structural integrity of that design, another KIWI-B1 was run on 1 September 1962 using liquid hydrogen, primarily to obtain data on potential two-phase flow problems. In that test, the liquid-hydrogen-cooled nozzles and hydrogen pump developed by Rocketdyne were used. The core failed and started ejecting fuel out of the nozzle as it increased in power level to about 900 MW. It was clearly not a suitable design. However, the test did provide significant information. It laid to rest the various apprehensions about two-phase hydrogen flow (27), the turbopump and nozzle in this first liquid hydrogen test operated well, and the control drums in the peripheral reflector effectively started up and controlled the reactor.

The next reactor to be tested was the KIWI-B4A design. This reactor was clearly the preferred design; it was expected to serve as the basis for the NERVA engine reactor design (28). It had the 19-hole hexagonal fuel element assembly configuration. Six of the uranium-fueled graphite fuel elements were placed

around the unfueled graphite center support element, and the entire module was supported by a hot end graphite cluster support block and a steel rod running through that central support system that was tied to a cold inlet support plate (29). The coolant passages in the fuel elements and the exterior surfaces at the hot end were coated with niobium carbide (NbC) to protect the fuel elements from corrosion. However, the high expectations for that test were quickly dimmed when the test started on 30 November 1962 and flashes of light in the nozzle exhaust indicated reactor core damage as the power was increased over about 250 MW. After disassembly, it was found that almost all of the fuel elements had broken (27) as a result of severe vibrations in the entire core.

Recovery from the KIWI-B4A Failure. In a meeting in the SNPO offices on 3 and 4 January 1963 with Dr. Norris Bradbury, the Director of Los Alamos; his deputy Dr. Raemer Schreiber, and some of their key people; the Deputy Manager and others from the SNPO; and representatives from NASA's Langley and Lewis Research Centers and the Marshall Space Flight Center, the author, Manager of SNPO, expressed his decision that there would be no further hot testing of a full reactor until thorough work was done to identify the causes of the problems that had occurred and to develop well-defined solutions to those problems. The Los Alamos Director objected saying that the Manager would kill the program if there were not continued reactor testing. The author responded that, on the contrary, the program would be killed if reactors continued to have major test failures. Dr. Bradbury then said that Los Alamos could make a "quick fix," but he did not define any specific fix in that meeting, nor subsequent to it. Any undefined quick fix was rejected.

The decision prevailed and a comprehensive collaborative program to understand and solve the problems was developed among the parties involved. It included reactor design analysis and extensive testing by Los Alamos and Westinghouse (30), including component and subsystem and vibrational tests. The Manager also decided that a cold flow, nonfueled and, therefore, nonfissioning test of KIWI-B4A should be conducted with sufficient instrumentation to confirm vibrations as the cause of the failures and to identify their sources. That test, run on 15 May 1963, confirmed a faulty design feature that resulted in interstitial flows between fuel elements and induced vibration and failure. Los Alamos had placed a peripheral seal at the inlet of the core so that the low exit pressure in the periphery permitted the interstitial or interfuel element flow to expand the core outward and induce element vibration. Westinghouse had actually been concerned about the need to bundle the core to limit the interstitial flow corrosion effects, so they had set the seal in its NERVA reactors at the core exit and were working on increased lateral support. Using the exit seal, the peripheral pressure was inward and increased the core bundling. (30, p. 398). The cold flow test clearly proved that the vibrations in KIWI-B4A were flow induced (28).

Based on those results and extensive analysis and component and section or "pie" testing at both Los Alamos and Westinghouse, changes in the peripheral seal and lateral support designs, similar to those that were being made in the KIWI and NERVA designs, were incorporated into a second cold flow test (KIWI-B4B-CF) in August 1963. No vibrations occurred in that cold flow reactor redesign. In October, based on those results, further supporting tests, and a

comprehensive review conducted at Los Alamos, approval was given to go forward with the redesign and with building and hot testing the KIWI-B4D reactor.

During the second half of 1963, NASA and the AEC were busy preparing budget requests for FY 1965. The Commission strongly favored requesting funds for flight testing the nuclear rocket RIFT system, even though full resolution of the KIWI-B4A problems had not yet been conclusively demonstrated. In a meeting with the Commission, the author proposed instead a comprehensive ground-based development program that would establish a sound technical basis for eventual commitment to flight systems and space missions. Although the Commission Chairman, Dr. Glenn T. Seaborg, and the NASA Administrator, James E. Webb, proposed the flight-test program, the President rejected it in their budget discussions. A revised budget request submitted by the Administrator and the Chairman to the President to cover the ground-based program was approved as the basis for the FY 1965 nuclear rocket budget of both agencies. As a result, the RIFT stage development was cancelled, but the NERVA development and further advanced work by Los Alamos on the higher power Phoebus reactor system was continued following the completion of the KIWI reactor tests in 1964.

The KIWI-B4D reactor was operated first as a cold flow unit in February 1964 without a problem. In May, the KIWI-B4D hot test was run. However, the test was cut short as a result of a hydrogen leak and resulting fire at the nozzle at the system's design power of 1000 MW. After disassembly, the reactor was in excellent shape and had no broken fuel elements. That test generated great euphoria throughout the program. That was followed by the KIWI-B4E tests in September, including the short restart (Fig. 10). Although some corrosion was apparent, the core held together, but it did have some broken fuel elements.

NERVA and Phoebus Testing. That KIWI-B4E test completed the Los Alamos KIWI reactor work, which Westinghouse and Aerojet General then extended to the NERVA reactor (NRX), engine breadboard (EST), and the full engine system (XE) tests and development. Los Alamos moved on to the Phoebus reactor work aimed at achieving higher power and temperatures (Fig. 8). Those Aerojet-Westinghouse NERVA tests continued to pave the way for an operational nuclear rocket engine, and the Los Alamos Phoebus work (Fig. 11) provided the information for larger engines. On 13 December 1967, the NRX-A6 achieved the program operating target of 60 minutes at 1100 MW (Fig. 8). On 11 June 1969, the full XE engine, shown in Fig. 12 in the downfiring Engine Test Stand with the full turbopump, nozzle, and propellant tank, ran successfully at 1100 MW and went through 28 starts from December 1968 to August 1969. The Phoebus reactor tests achieved more than 4000 MW, equivalent to more than 200,000 pounds of thrust.

It is important to recognize that, in addition to overcoming the early structural problems discussed before, the tests conducted from 1964 through 1969 applied continued improvements in fuel material and protection from hydrogen corrosion (31). They all had fuel elements containing small pyrocarbon-coated uranium carbide spheres in the graphite matrix and used niobium or zirconium carbide to protect the carbon in the matrix from hydrogen corrosion. Work to evaluate further fuel element improvements was conducted by Los Alamos in the



Figure 10. KIWI-B4E at the test cell.

Nuclear Furnace fuel element tests, which ran for 109 minutes at high power densities and temperatures in a radiative environment. Those tests also included a scrubber to remove fission products from the exhaust.

Conclusion of Nuclear Rocket Rover Program. In spite of the view in NASA that nuclear propulsion using the NERVA engine would be needed to resume lunar exploration in the 1980s and other advanced missions, its proposed FY 1972 budget continued the “engine development at a minimum rate—due to fiscal constraints” (32). However, in the face of FY 1972 space program budget ceilings set by the Office of Management and Budget (OMB), NASA offered termination of the NERVA engine development program although it was “the only large scale advanced propulsion system underway.” The program was killed. As pointed out in Reference 33, “This decision ended a longstanding NASA policy of developing advanced engines well before there was need for them.”

Based on the progress that had been made in advancing U.S. nuclear rocket technology in the program, there is no question that these results indicate that technology is available that is suitable for application in high-payload, deep space missions when such missions are established as space objectives. However, such missions have not yet been defined in spite of continued interest over decades in ultimate human missions to Mars. The headline in the Wall Street



Figure 11. Phoebus 1B being moved to the test cell.

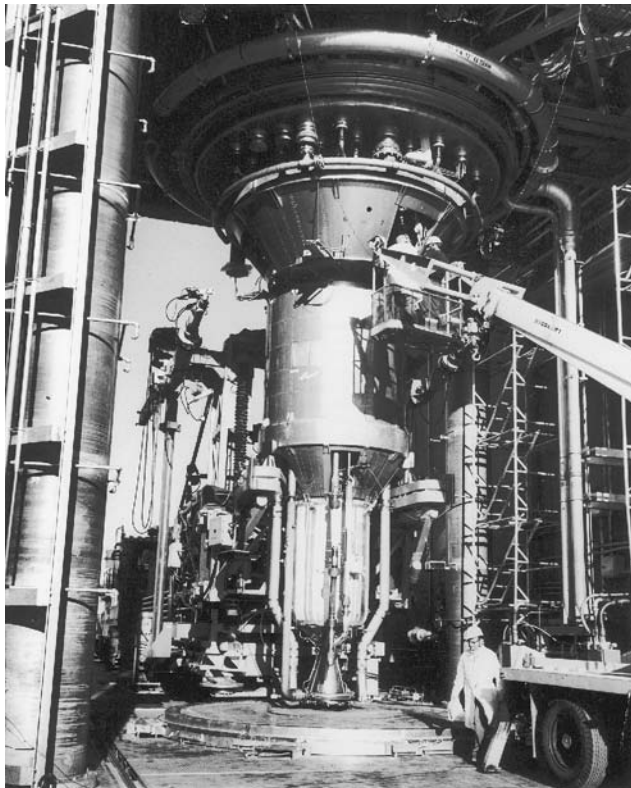


Figure 12. The XE Engine in the engine test stand for full engine test.

Journal on 11 June 1964 after the successful KIWI-B4D test recognized that "Nuclear Rocket's Technical Problems Seem Solved, but Space Mission for it is Lacking." T.A. Heppenheimer makes the same point in his recent book "The Space Shuttle Decision" (33) when, referring to the progress made after the KIWI-B4A reactor failure, he points out that "The rapid pace of advances in Nevada contrasted painfully with the lack of plans in Washington. With NASA having no approved post-Apollo future, it was quite possible to anticipate a time when Aerojet might build a well-tested NERVA, ready for flight, only to find that NASA had no reason to use it."

However, some analysis, testing, and reporting on the capabilities of nuclear rocket propulsion based on the advances provided by the NERVA and Phoebus work and other nuclear rocket concepts for deep space missions has continued. Especially as NASA continues to examine the possibility of Mars missions extending to human exploration, that past work continues to be currently relevant. Reference 34 is an excellent example of ongoing nuclear rocket activities and examination as are continued conferences conducted in Russia, discussed later. The so-called Tom Stafford report (21) further emphasizes this point indicating, "Advanced nuclear propulsion techniques can shorten the transit time, provide flexible surface stay time, significantly reduce the propellant mass to low Earth orbit and increase the available launch opportunity." The interest in Mars exploration is great, the ultimate need for nuclear propulsion for such human exploration is apparent, the technology is available, but the commitment to such missions is still remote today. A serious question is how the knowledge base established by the nuclear rocket program can be retained until the time when such a mission commitment is made.

Russian Nuclear Rocket Development Activities. As indicated before, Russia continues to hold conferences devoted to space nuclear systems. The first of these open conferences was held in 1990 in Obninsk. The tenth conference was held in 2000, again, in Obninsk. Those conferences have included participants from the major organizations and activities involved in space nuclear system development in the Commonwealth of Independent States (CIS) and the United States. As a follow-up to attendance at the Third Specialist Conference on "*Nuclear Power Engineering in Space: Nuclear Rocket Engines*," held in September 1992 in Semipalatinsk-21 in the Republic of Kazakhstan, the U.S. Department of Energy organized a team to visit the many Institutes in the CIS that have been involved with space nuclear power and propulsion research and development in addition to the Russian Ministry of Atomic Energy and the then relatively new Space Agency. That team was led by the Director of the Office of Space and Defense Power Systems in DOE and included representatives from the DOE and its laboratories, from the DOD including its Phillips Laboratory and its Strategic Defense Initiative Organization, and consultants that have experience in the areas being examined. The author was part of that team. A principal purpose of these visits was to familiarize the DOE and DOD with Russian capabilities that might suggest cooperative activities in areas considered important to the United States. Such a cooperative space effort had been suggested by the Minister of Atomic Energy. Much of the information concerning the Russian nuclear rocket programs presented here is drawn from the discussions of that team at the conference and in visits with the Institutes, which were fully reported in

Reference 35. That report concludes that there are many areas where cooperative efforts could be beneficial; some joint work has been conducted.

The history of the Soviet work in nuclear energy was summarized at the start of the Conference by its Chairman, Dr. N.N. Ponomarev-Stepnoi of the Kurchatov Institute in Moscow. He pointed out that 4 years after controlled fission was achieved in Chicago, the Kurchatov Institute achieved its first fission in a still operating nuclear pile. Their first work was aimed at research and plutonium production, but later they went on to power reactors and then ship and aircraft propulsion. No nuclear aircraft were ever flown. In his comments on nuclear rocket propulsion safety, the Chairman emphasized his view that sub-criticality must be ensured before the rocket stage achieves its high operational orbit. This was similar to the view NASA expressed from the start of its role in Rover and was specified for the NERVA design.

Turning to development of nuclear rocket reactors, he and the following speaker from the Institute of Scientific Industrial Association (SIA) "Luch" described the nearby reactor test facilities that were later visited by the DOE Team. The Impulse Graphite Reactor (IGR) (similar to the U.S. TREAT – Transient Reactor Test facility) went critical in June 1960, and the first nuclear thermal propulsion fuel tests were started there in 1964 when gaseous hydrogen was made available for tests. Another reactor built at the Baikal site (64 km away) was the IVG-1, which could test up to seven clusters of fuel assemblies in the central test section using gaseous hydrogen. The first tests in that reactor were started in 1975, and 30 hot firings were conducted up to 1985. A second reactor, the IRGIT, was built at that site to test more typical rocket reactor cores or fuel clusters. It became operational in 1978 but after only four test series were run, its pressure vessel failed due to hydrogen embrittlement. All of this indicates that the former Soviet Union program on nuclear rocket propulsion started well after the U.S. program. It is also apparent that the available reactor test facilities have no liquid hydrogen capability, no reactor testing was done there using liquid hydrogen, and no full reactor or engine tests were ever conducted on nuclear rocket propulsion. However, substantial effort was devoted to and progress was made on materials development suitable for fuel elements and structures of nuclear rocket reactors. Some work continues in those areas.

In the conference and in its meetings with eight Institutes, it became apparent to the Team that the former Soviet Union had established a large, competitive, and overlapping infrastructure for work on space nuclear systems. It was also apparent that some of this capability had been established for work in developing nuclear weapons. Though the start of the Soviet work followed U.S. activity, much of it continued well after the U.S. nuclear rocket program ended in 1973. The indication that the Soviet program knew of the U.S. activities was apparent in much of the work that was presented during the various visits, but it was also emphasized when the Director of the Research and Development Institute of Power Engineering (RDIPE) said that they knew of the U.S. work even though no one had given them the SNPO reports. That Director of RDIPE later became the Minister of Minatom, the Ministry of Atomic Energy. There is no question that Russia has had, and continues to have, interest in nuclear rocket propulsion for use when potential space missions require it.

The Nuclear Ramjet Pluto Program

As pointed out before, the Lawrence Livermore Radiation Laboratory (LRL) was assigned to work on the nuclear ramjet program in 1957. The Marquardt and later, the Chance-Vought companies were involved in the program with LRL under Atomic Energy Commission and Air Force contracts; Atomics International provided some fuel development support. Several different names were used for the program. The development work at Livermore was referred to as the Pluto project, and their ramjet reactors to be tested were named Tory reactors. The overall nuclear ramjet missile system was the Supersonic Low Altitude Missile (SLAM).

A schematic drawing of the nuclear ramjet is shown in Fig. 2, which identifies its principal components. As in the nuclear rocket program, major emphasis was required and placed on selecting and developing high-temperature materials suitable for the reactor system. However, in the ramjet, they would have to be compatible with the oxidizing atmosphere of the air working fluid. In addition, keeping the reactor small and light was also required to ensure that the ramjet missile delivery system could carry a significant missile payload. The missile speed was expected to be at least twice the speed of sound as it flew its low-altitude, low-detection trajectory. In fact, NACA research had indicated that wingless conical or circular ramjet vehicles could achieve reasonable lift/drag ratios at Mach numbers between 2 and 4; above Mach 5, they are equal to the best winged vehicles (36). However, as also discussed in the Introduction to this article, such speeds introduce significant design requirements and development work on the inlet system, and they lead to high pressure drops and stresses in the reactor, in addition to those from temperature variations. The nuclear ramjet design aimed at speeds of three times the speed of sound (19).

Dr. Theodore Merkle, the Director of the Pluto program and Associate Director at Livermore, described the system and discussed the very important reactor material considerations and their selection in the hearings of Reference 12 and in Reference 11. The strength of materials at high temperature and resistance to oxygen were the key issues in material selection for the nuclear ramjet. For example, some of the high-temperature materials such as graphite and tungsten and carbides such as zirconium and tantalum carbide, which were considered and some applied in the rocket program, could not be used in the ramjet because they burn up in oxygen. The conclusion was that certain ceramics would best meet the reactor requirements; this led to a mixture of uranium oxide (UO_2) in a beryllium oxide moderator as the preferred fuel element material (38).

Preparing for tests after designation as the Pluto development organization in 1957 involved major facility construction at the Nevada Test Site (NTS) in an area next to the Nuclear Rocket Development Station (NRDS) at Jackass Flats (36). For example, in addition to the test stand and the disassembly building, a hot nuclear critical assembly building was built and a very high capacity source of high pressure air—a million pounds—was required to simulate the ramjet flight conditions and to serve as the reactor coolant and thrust fluid of the system (37). That pressurized air system required 25 miles of oil-well casing and used compressors from the Navy's Groton submarine base to supply the high-pressure air that also was heated to simulate the system's supersonic flight. In addition to

the extensive facility work, LRL was heavily involved in its design, criticality testing, and fuel materials development, to arrive at its test reactor designs.

The first test reactor, the Tory II-A1, was a small, low-power test reactor. The main purpose of testing it was to verify the predicted integrity of the reactor core materials and to study the aerothermodynamic behavior of the core-air heat exchange system under conditions that simulated low-altitude supersonic flight (19). General Electric's Aircraft Nuclear Propulsion Department fabricated the fuel elements for Tory II-A1. The graphite peripheral reflector was water cooled, and control rods were located in the reflector. Sections at the inlet and outlet ends of the reactor made of unfueled beryllium oxide also acted as neutron reflectors (39). The reactor was assembled at LRL and went critical there on 7 October 1960. It was then moved to the NTS where it was initially tested at 40 MW, about a third of its design power, on 14 May 1961 for about 60 seconds at an inlet air temperature of 400°F and a maximum fuel element temperature of 2250°F (19). The results exceeded expectations. Further test runs were successfully made at full power in September and October 1961 (40). The delay in full power testing resulted from the need to replace the air ducting in the test bunker with high-temperature duct alloy. Although the planned reactor test schedule (21) called for further tests of a second backup Tory II-A reactor, the results of the Tory II-A1 were sufficient to eliminate the need for that second, small, low-power test system.

The program was then directed toward the development and testing of the Tory II-C flight type reactor system. Its configuration differed significantly from the Tory II-A. Its control system was contained within the core; it had a thin air-cooled reflector and improved fuel elements. Its fuel was made by Coors Porcelain Co. (38). That Tory II-C reactor was assembled and went critical at Livermore in July 1963; criticality testing continued through September. After shipment to the test site (NTS), facility welding deficiencies were found that delayed testing of the reactor till 20 May 1964 (40).

Conclusion of the Nuclear Ramjet Pluto Program. Though the Tory II-C test in May was fully successful, the Department of Defense Director of Defense Research and Engineering (DDR&E), Dr. Harold Brown (a former Director of the Livermore Laboratory), notified the Air Force on 6 July 1964 that the Low-Altitude Supersonic Vehicle (LASV) program would be cancelled by DOD, and he notified the AEC that the DOD would not support a flight test of the nuclear ramjet. For several years, in spite of Air Force and Navy encouragement of nuclear ramjet development, the program had been going through major questioning in congressional hearings and in the DOD concerning its anticipated application to a military mission. In fact, the House Appropriations Committee had almost eliminated the AEC FY 1965 funds for the program and threatened to cut the DOD funds if the DOD did not plan to move forward to develop the vehicle system required to flight test the reactor that had been successfully tested. Based on the success of the Tory II-C reactor testing, some argued that the next logical step was to prepare for a flight test of the system to prove its readiness for mission application. However, many considered the cost of such a flight program excessive in the face of no clear need or application for the system. The DDR&E took that position, as did Secretary of Defense McNamara, who indicated that the chances of deploying the system were slight (41). As a result,

the program was cancelled in spite of the sound technical progress that demonstrated the engine feasibility of the system.

The front page headline of the Smithsonian's April/May 1990 issue of *Air & Space* referring to Gregg Harken's article (37), which in some places expressed concerns that were considered by some to be exaggerated and even erroneous, summed up the conclusions of many: "Project Pluto: How America almost built a nightmare missile." The article was entitled "The Flying Crowbar" which was the description of the nuclear ramjet system used by Dr. Merkle in his testimony in Reference 12 to indicate its structural solidity. Fundamentally, the existing reactor technology could be further advanced, but in view of the already existing and adequate chemical rocket ballistic missile capability, the lack of any mission requirement eliminated that need. There are no missions foreseen for the nuclear ramjet.

BIBLIOGRAPHY

1. *The Papers of Wilbur and Orville Wright*, Vol. One: 1899–1905. M.W. McFarland (ed.). McGraw-Hill, 1953.
2. *The Papers of Robert H. Goddard*, Vol. II: 1925–1937. E.C. Goddard (ed.). McGraw-Hill.
3. *Controlled Nuclear Chain Reaction—The First 50 Years*, published with support from The University of Chicago Board of Governors for Argonne National Laboratory. American Nuclear Society, LaGrange Park, IL, 1992.
4. Gosling, F.G. *The Manhattan Project—Making the Atomic Bomb*. U.S. Department of Energy, Energy History Series, DOE/HR-0096, Washington, DC, 1994.
5. Cleator, P.E. *Rockets Through Space*. Simon and Schuster, New York, 1936.
6. *The Papers of Robert H. Goddard*, Vol. III: 1938–1945.
7. President John F. Kennedy's Special Message to the Congress on Urgent National Needs, 25 May 1961. Public Papers of the Presidents of the United States, John F. Kennedy, 20 January to 31 December 1961. U.S. Government Printing Office, Washington, DC, 1962.
8. Dawson, V.P. *Engines and Innovation*. Lewis Laboratory and American Propulsion Technology. The NASA History Series, Washington, DC, 1991.
9. Hewlett, R.G., and O.E. Anderson, Jr. *The New World, 1936/1946*, Vol. I, *A History of the United States Atomic Energy Commission*. Pennsylvania State University Press, University Park, Pennsylvania, 1962.
10. NACA Conference on Ram Jets. A compilation of papers presented by NACA Staff Members, Aircraft Engine Research Laboratory, Cleveland, OH, October 29, 1946; *NACA Conference on Aircraft Propulsion Systems Research*. Lewis Flight Propulsion Laboratory, Cleveland, OH, January 1950; *NACA Conference on Turbojet Engines for Supersonic Propulsion*. A compilation of technical material presented by Lewis Flight Propulsion Laboratory, October 1953; *NACA 1957 Flight Propulsion Conference*. Lewis Flight Propulsion Laboratory (all at NASA Library, Washington, DC.).
11. Gantz, K.F. *USAF Nuclear Flight. United States Air Force Programs for Atomic Jets, Missiles, and Rockets*. Duell, Sloan and Pearce, New York,
12. *Outer Space Propulsion by Nuclear Energy*. Hearings before Subcommittees of the Joint Committee on Atomic Energy. Eighty-Fifth Congress of the United States. 22, 23 January and 6 February 1958, pp. 3–6 and p. 209.
13. Sloop, J.L. *Liquid Hydrogen As A Propulsion Fuel, 1945–1959*. NASA History Series, NASA SP-4404. National Aeronautics and Space Administration, Washington, DC, 1978.

14. The Birth of NASA. The Diary of T. Keith Glennan, The NASA History Series, NASA SP-4105. National Aeronautics and Space Administration, Washington, DC, 1993.
15. Hewlett, R.G. and F. Duncan. Atomic Shield, 1947/1952, Vol. II. A History of the United States Atomic Energy Commission. Washington, DC, 1969.
16. Green, H.P., and A. Rosenthal. *Government of the Atom. A Study Sponsored by the National Law Center of the George Washington University*. Atherton Press, New York, 1963.
17. Memorandum of Understanding, August 1960, Exhibit A to AEC–NASA Interagency Agreement on Rover Program, NASA General Management Instruction 2-3-17. NASA History Office, Washington, DC, 1960; AEC and NASA News Release No. 60-252, 31 August 1960, Joint AEC–NASA Nuclear Propulsion Office Established. NASA History Office, Washington, DC.
18. Project Rover (U.S. Nuclear Rocket Development Program. Hearings before the Committee on Science and Astronautics U.S. House of Representatives, February and March 1961; NASA Scientific and Technical Programs. Hearings before the Committee on Aeronautical and Space Sciences United States Senate, 1 March 1961.
19. Nuclear Energy for Space Propulsion and Auxiliary Power. Hearings before the Subcommittee on Research, Development, and Radiation of the Joint Committee on Atomic Energy, 28 August 1961.
20. Bilstein, R.E. Stages to Saturn. A Technological History of the Apollo/Saturn Launch Vehicles, NASA SP-4206, The NASA History Series. NASA History Office, Washington, DC, 1996.
21. America at the Threshold: Report of the Synthesis Group on America's Space Exploration Initiative, Thomas P. Stafford, Chairman, Submitted to the Chairman, National Space Council, 3 May 1991.
22. Bussard, R.W., and R.D. DeLauer. *Los Alamos Scientific Laboratory, University of California: Nuclear Rocket Propulsion*. McGraw-Hill, New York, Toronto, London, 1958; Bussard, R.A. Fundamentals of nuclear propulsion. In Willaume, R.A., A. Jaumotte, and R.A. Bussard. *Nuclear Thermal and Electric Rocket Propulsion – Fundamentals, Systems and Applications*, AGARDograph 101, AGARD/NATO First and Second Lecture Series of 1962 and 1964 in Cooperation with the Universite Libre de Bruxelles. Gordon and Breach, New York, 1967.
23. Rom, F.E., E.W. Sams, and R.E. Hyland. 2. Nuclear Rockets. *NACA 1957 Flight Propulsion Conference*. Vol. II 22 November 1957, Lewis Flight Propulsion Laboratory, Cleveland, OH. NASA Library, Washington, DC; Rom, F.E. NASA Lewis Research Center: Fast and moderated reactors and applications of low-power nuclear rockets, Willaume, R.A., A. Jaumotte, and R.W. Bussard, *Nuclear Thermal and Electric Propulsion*, AGARD/NATO First and Second Lecture Series of 1962 and 1964. Gordon and Breach, New York, 1967.
24. MacMillan, D.P. Los Alamos Scientific Laboratory: High-temperature materials for rocket reactors. *Nucleonics* 19 (4): 1961.
25. Schreiber, R.E. Kiwi tests pave way to Rover. *Nucleonics* 19 (4): (1961).
26. Dewar, J.A. *To the End of the Solar System: The Nuclear Powered Rocket*. Book being reviewed for possible publication.
27. Spence, R.W. The Rover nuclear rocket program. *Science* 160 (3831): (1968).
28. Finger, H.B., J. Lazar, and J.J. Lynch. A survey of space nuclear propulsion. National Aeronautics and Space Administration, Atomic Energy Commission, Washington, D.C. *Presentation to 1st AIAA Annual Meeting*, 30 June 1964.
29. Klein, M. *Nuclear Rocket Technology Conference*. Lewis Research Center Cleveland, OH, 1966. National Aeronautics and Space Administration, Washington, DC.
30. Simpson, J.W. *Nuclear Power from Underseas to Outer Space*. American Nuclear Society, LaGrange Park, IL, 1995.

31. Lyon, L.L. Performance of (U,Zr)C-graphite (composite) and of (U,Zr)C (carbide) fuel elements in the nuclear furnace 1 test reactor. LA-5398-MS Informal Report UC-33, September, 1973. Los Alamos Scientific Laboratory, Los Alamos, NM.
32. Letter from G.M. Low, Acting Administrator, National Aeronautics and Space Administration to Honorable G.P. Schultz, Director, Office of Management and Budget, 30 September 1970.
33. Heppenheimer, T.A. The Space Shuttle Decision. NASA's Search for a Reusable Space Vehicle, NASA History Series. NASA Sp-4221. NASA History Office, Washington, DC, 1999.
34. Nuclear Thermal Propulsion. A Joint NASA/DOE/DOD Workshop. NASA Conference Publication 10079, *Proc. Nuclear Thermal Propulsion Workshop* sponsored by NASA Lewis Research Center, Cleveland, OH, July 10–12, 1990.
35. An Examination of the Space Nuclear Power and Propulsion Activities of the Commonwealth of Independent States. Report of the U.S. DOE Space Nuclear Power and Propulsion Team, 20 September–6 October 1992, U.S. Department of Energy, Washington, DC.
36. Butz, J.S., Jr. Lack of Engineering Data Delays Nuclear Ramjet. *Aviation Week* (March 2, 1959).
37. Herken, G. The Flying Crowbar. *Air & Space*, Smithsonian (April/May 1990).
38. Hawkes, R. Tory II-A nuclear ramjet nears test. *Aviation Week* (December 19, 1960).
39. Space Nuclear Power Applications. Hearings before the Subcommittee on Research, Development, and Radiation of the Joint Committee on Atomic Energy. September 19, 1962, 87th Congress of the United States, Second Session on Space Nuclear Power Applications.
40. AEC Authorizing Legislation Fiscal Year 1965. Hearings before the Joint Committee on Atomic Energy. Part 2. February 20, 1964, 88th Congress of the United States.
41. Trainor, J. DOD Decides to Cancel LASV. *Missiles and Rockets* (July 13, 1964) and *Missiles and Rockets* (July 20, 1964).

HAROLD B. FINGER

formerly with National Aeronautics and
Space Administration and Atomic
Energy Commission, Washington, D.C.

O

OPTICAL ASTROMETRY FROM SPACE

Astrometry plays a very particular part in the realm of astronomy. On the one hand, it is essentially an ensemble of techniques that provides some essential data to astronomers and astrophysicists about celestial objects. On the other hand, until the second half of the nineteenth century, what is now called astrometric observations were the only astronomical activity that existed. Actually, astronomy has a tradition that goes back to Egyptian, Assyrian, and Greek astronomy. Astrometry is the oldest of all sciences and still is a scientific domain of its own, encompassed by theoretical developments such as stellar dynamics and celestial mechanics now supported by the theory of general relativity.

One can define astrometry as the part of astronomy that measures the apparent positions of celestial bodies on the sky. And, because these positions vary with time, the objective is to describe and study these motions that, for stars, provide two essential parameters: the proper motion and the parallax from which the distance is derived. As an extension, one ascribes also to astrometry the measurement of apparent dimensions and shapes of celestial bodies. However, in this article, we shall consider the determination of star positions, the primary goal of space astrometry.

The physical quantities that are measured by astrometry are angles that are often very small. Radians are not used in astrometry; the basic units are degrees and seconds of arc (denoted"). Smaller units are necessary, and astrometrists are currently using milliseconds of arc (denoted mas, that is, a milli-arcsecond) and now are starting to use one millionth part of a second of arc (μ as). Their respective values are close to 5×10^{-9} and 5×10^{-12} radians.

Essentials of Astrometry

Before discussing how astrometric measurements are performed, it is appropriate to present some basics that have to be known for further understanding.

Reference Systems and Frames. The position of a point in the sky is defined by its two spherical coordinates. The most frequently used are the equatorial system. The principal plane is the celestial equator coplanar with Earth's equator. Starting from the vernal equinox at the intersection of the equator and the ecliptic, the right ascensions (denoted α) are reckoned counterclockwise. The second angular coordinate is the declination (δ), counted from the equator, positive to the North, negative to the South (Fig. 1). We shall also use the ecliptic system. The principal plane is the ecliptic, and the celestial longitudes (λ) are reckoned counterclockwise from the vernal equinox. The second coordinate is the celestial latitude (β), also shown in Fig. 1. Because both the physical equator and ecliptic are moving, the principal planes have a conventional fixed position (sometimes called mean equator or ecliptic).

These coordinate or reference systems are virtual and are obviously not actually located in the sky. A reference system is actually determined by assigning a consistent set of coordinates to a number of objects (fiducial points). Such a catalog of positions is said to be a reference frame. The position of any object is deduced from relative measurements with respect to fiducial points. Another important condition is that the coordinate systems must be fixed in time, so that the apparent motions of celestial bodies are not falsified by spurious rotation. This is realized now by choosing as fiducial points very distant objects

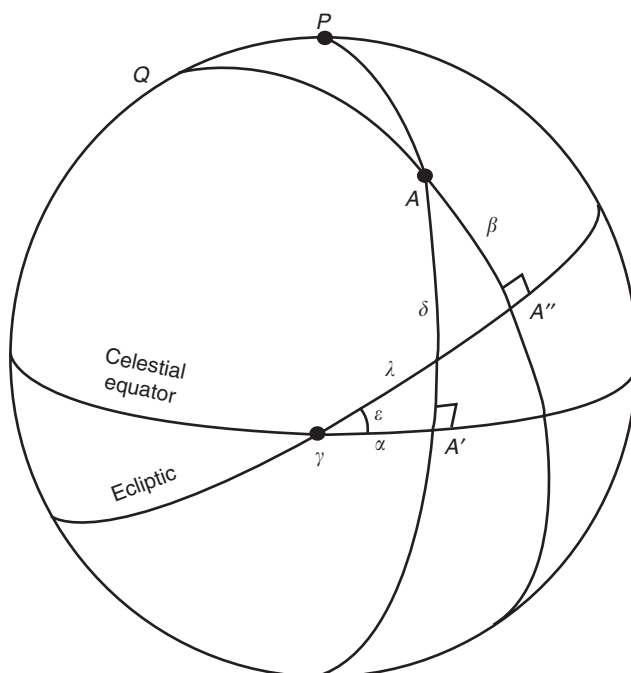


Figure 1. Equatorial (α , δ , pole P) and ecliptic (λ , β , pole Q) coordinate systems.

(quasars or galaxies) whose motions are slower than the speed of light and so appear negligible as seen from Earth. The system so defined is the International Celestial Reference System (ICRS) (1), and the catalog of fiducial extragalactic objects is the International Celestial Reference Frame (ICRF), a radio-source catalog extended (2) to optical wavelengths by the Hipparcos catalog (see later).

Apparent and True Positions. The direction from which the light arrives at the instrument has undergone a series of deviations. For this reason, it arrives from an apparent position of the star, not the true one. There are three causes for this deviation.

Atmospheric Refraction. Light from outer space is bent progressively as it enters the atmosphere which is composed of layers of different refractive indexes. The integrated effect depends upon the pressure, temperature, and humidity of the atmosphere and the wavelength of the light. The lower the object in the sky, the larger and the more uncertain the correction to be applied. In practice, ground-based astrometric observations are not performed below 60° zenith distance.

Aberration. The apparent direction of a source is a combination of the direction from which the light arrives and the the velocity of the observer. In ground-based observations, one distinguishes the diurnal aberration due to the motion of the observer as a consequence of Earth's rotation and stellar aberration due to the motion of Earth around the Sun. In astrometry from space, the diurnal aberration is replaced by the orbital aberration due to the motion of the satellite in its orbit.

In the Newtonian approach, the apparent direction \mathbf{r}' of a star is linked to the undeviated direction \mathbf{r} by

$$\mathbf{r} = \mathbf{r}' - r\mathbf{V}/c, \quad (1)$$

where \mathbf{V} is the velocity of the observer and c is the speed of light. For very precise astrometry, one must use a more complex formulation based on the theory of general relativity (3).

Relativistic Light Deflection. Following the theory of general relativity, a massive body produces a curvature of space, and the geodesic followed by the light ray is not a straight line: it deviates by a small amount toward the massive body. For the Sun, the deviation is

$$\gamma = 0.00407'' \cotan \theta/2, \quad (2)$$

where θ is the angle between the directions of the star and of the Sun.

Parallactic Displacements. The true position obtained after applying the corrections described above refers to a moving observing site. For positional comparisons, this is not convenient, and it is necessary to refer to a more stable origin of the coordinate system. The correction to be applied to get the direction viewed from another point is the parallactic displacement or correction. Two cases are useful.

Geocentric Coordinates. The parallactic correction necessary to shift from ground-based or satellite-based observations to the center of Earth is totally

negligible for a star. This is not the case for observations of objects in the solar system.

Barycentric Coordinates. This coordinate system is centered at the barycenter of the solar system. It is the only point whose motion in space is linear with very high accuracy, because it corresponds to an orbit around the center of the Galaxy described in 280 million years. For all practical applications, it can indeed be considered linear without any dynamic effect on the members of the solar system. The construction of the parallactic correction is sketched in Fig. 2. Let B be the barycenter of the solar system, E the center of Earth in its orbit C around B . Let S be the actual position of a star and r its distance; $\mathbf{r} = \mathbf{BS}$. From Earth, S is seen along the vector $\mathbf{r}' = \mathbf{ES}$. The apparent direction \mathbf{ES} differs from the barycentric direction \mathbf{BS} by the angle

$$p = (\mathbf{r}, \mathbf{r}') = (\mathbf{ES}, \mathbf{BS}). \quad (3)$$

If we call θ the angle $(\mathbf{EB}, \mathbf{BS})$,

$$\sin p = (R/r) \sin \theta, \quad (4)$$

where R is the length BE . So the variation of the parallactic displacement p with time is a function of the motion of Earth on C , usually taken as an ellipse but may be made more precise using ephemerides.

Stellar Parallax. The angle p is of the order of or smaller than R/r . The convention is to express R/r not in radians, but in seconds of arc and define a

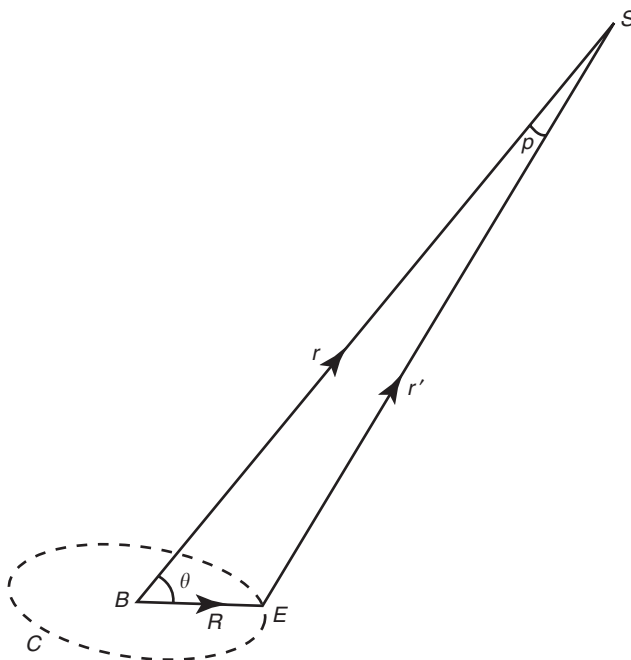


Figure 2. Stellar parallactic correction.

quantity called stellar parallax, or simply parallax, which is equal to the angle p when R is equal to the mean Earth–Sun distance, that is one astronomical unit (149,597,870 km). Because $1''$ is equal to $2\pi/(360 \times 3600)$ radians, the distance for which the parallax is equal to $1''$ is 206, 265 AU or 3.2616 light-years. This distance is called the *parsec* (parallax-second, abbreviated as ps) and is the commonly used distance unit outside the solar system. With this unit, the distance r is simply

$$r = 1/\varpi, \quad (5)$$

where ϖ is the parallax p expressed in seconds of arc. Note that the nearest star, Proxima Centauri, has a parallax of $0.762''$. Very few stars are at distances smaller than 10 ps and most of the stars of astrophysical interest have parallaxes of the order of a few hundredths or even thousandths of an arc second. This implies that to be significant, their parallaxes should be determined at least at an accuracy of 1 mas, and even much more. This is the major challenge to astrometry nowadays, and this is the main driver for very accurate astrometrical measurements, possible only from space.

Motion of Stars. Stars move in space, and observing their apparent motion in the sky allows us to access dynamic properties of groups of stars (double and multiple stars, star clusters and the Galaxy itself). Two types of motion can be distinguished.

Proper Motions. The position of a star with respect to a fixed celestial reference frame varies with time. Very often, the motion is linear and is expressed as yearly variations of the coordinates, called proper motion:

$$\mu_\delta = d\delta/dt; \quad (6)$$

$$\mu_\alpha = d\alpha/dt. \quad (7)$$

In local coordinates centered at the star, the components of proper motion are $\mu_\alpha \cos \delta$ and μ_δ . It is often useful to express the tangential velocity in kilometers per second. This is possible only if the distance is known and, after some transformations of units, for any component μ of the proper motion,

$$V = \mu/4.74\varpi \text{ km s}^{-1}. \quad (8)$$

This is the projection of the actual velocity of the star on the plane perpendicular to the direction of the star. The third component of space velocity is radial velocity, which is measured by spectroscopic techniques. It is obtained by measuring the Doppler shift $\Delta\lambda$ of spectral lines at a wavelength λ_0 :

$$V_R = c\Delta\lambda/\lambda_0, \quad (9)$$

where c is the speed of light.

Sometimes, the path of a star is not linear. This means that it is attracted by some invisible body, generally a companion of the star such as another faint star, a brown dwarf, or a planet. The star is then called an astrometric double star.

Relative Motions. One star moves with respect to another one, close to it. The observation of such motions is of particular importance in the case of double stars, when one of them revolves around the other following Newton's law of universal gravitation. If M_1 and M_2 are the masses of the components, the force that attracts M_2 by M_1 is

$$\mathbf{F} = \frac{kM_1M_2}{|\boldsymbol{\rho}|^3} \boldsymbol{\rho} \quad (10)$$

where k is the gravitational constant and $\boldsymbol{\rho}$ is the radius vector between the components. Observing double stars is one of the main activities in astrometry. If the distance to the star is known, the sum of their masses can be determined by modeling the apparent path as the projection of a Keplerian orbit. In addition, simultaneous knowledge of the radial velocities or the actual absolute path of both components, as shown in Fig. 3, taken from Reference 4, also allows us to determine M_1/M_2 and hence obtain the values of both masses (5).

The determination of relative motions in a star cluster is the material from which one can study the kinematic and dynamic properties of the cluster and compare the results to models. An example of what can be achieved with the best presently available astrometric data and radial velocities is found in Reference 6.

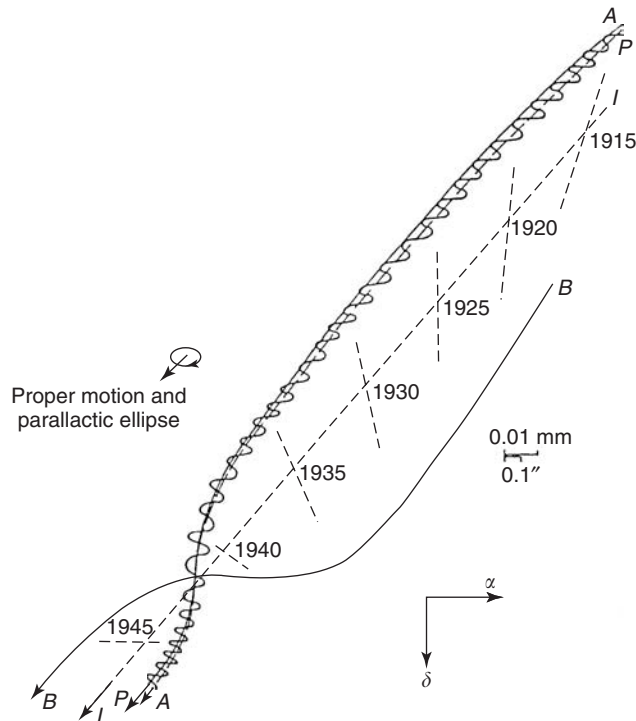


Figure 3. Geocentric path of the components of a double star (99 Herculis).

Ground-Based Astrometry

Before describing what space has brought to astrometry and what it is expected to bring in the future, it is necessary, for comparison, to present the achievements of ground-based astrometric observations. One must distinguish two classes of instruments that differ by the area of the sky measured. Detailed descriptions of the instruments mentioned in this section can be found in Reference 7.

Small-Field Astrometry. The positions of the celestial bodies are in this case measured with respect to neighboring stars instrument's in the field of view. The typical small-field instrument is the telescope that has either a photographic plate or—more generally now—charge-coupled device (CCD) arrays at its focal plane. The field of view of an array is a fraction of a square degree, but there is way to increase it by the scan mode. In this observing method, the telescope is fixed, and the charge transfer in the CCD is continuous at the rate of the diurnal motion. By this technique, it is possible to scan a narrow but long band along a declination circle.

The field of view of photographic plates, depends on the focal distance of the telescope. In Schmidt telescopes, it is as large as $5^\circ \times 5^\circ$, and the precision of position measurements is limited to $0.2''$. In long-focus telescopes (10 to 20 meters), the field is reduced to a fraction of a square degree, but the relative position of a star with respect to another is as close as a few hundredths of a second of arc. By combining several tens of long-focus observations, one obtains the best ground-based parallaxes to a few mas precision (8). Michelson interferometry that has a coherent field of a few seconds of arc is used to measure star dimensions and very close double stars at accuracies of the order of one mas (9). Speckle interferometry which allows a larger field, is perfect for double star observations, and reaches precision of a few mas (10).

Semiglobal Astrometry. Instruments in this category are designed to determine relative positions of widely separated celestial bodies, much farther apart than their fields of view. However, because one cannot see the whole sky from any place on Earth, one is constrained to some regions; the corresponding astrometry is called semiglobal, rather than global.

The oldest, and still the most used, instrument of this kind is the transit. It consists of a refractor telescope that can rotate around an east–west axis (6). The optical axis can be set to any direction on the local meridian. A micrometer registers the path of the star image on the focal plane and, by interpolation, one gets the time t of the transit of the star through the meridian. Then t is transformed into T , the Greenwich sidereal time and one obtains the right ascension of the star by

$$\alpha = L + T, \quad (11)$$

where L is the longitude of the observatory. Simultaneously, the inclination θ of the axis of the tube is measured using a divided circle, and one gets the declination δ by a formula such as

$$\delta = \phi + \theta - 90^\circ \quad (12)$$

for a southern transit in the northern hemisphere, or similar formulas in other configurations, where ϕ is the latitude of the observatory. After determining all of the instrumental parameters, one obtains precisions of the order of $0.1''$ or slightly better for stars that can be as much as 120° apart in observations that last the whole night.

Astrolabes have been also used for semiglobal astrometry. They observe star transits through a horizontal celestial small circle a little more accurately, but they are much less efficient (6). Recently, Michelson interferometry has been tested to determine relative star positions in various directions (11). Although there are hopes that this technique can give much more accurate results, the only existing instrument (the Navy Prototype Optical Interferometer in Arizona) is too new to permit definite statements on its performance. In any case, the number of observations per night remains quite limited in comparison with a transit instrument which may observe several stars during a single night.

Limitations of Ground-Based Astrometry. Except for limited instances in long-focus or interferometric small-field astrometry, the actual uncertainties in observations are of the order of $0.1''$. Even if the same stars are reobserved many times, the resulting precision is hardly improved by more than a factor of 3 due to the presence of systematic errors. Compared with the milliseconds of arc required for astrophysically significant results, at least one order of magnitude is to be gained. Several reasons exist for this fundamental limitation. Let us examine them.

Atmospheric Refraction. As already mentioned, refraction is not fully predictable. It varies with time and position in the sky, and the correction applied is not perfectly modeled. The result is that, in semiglobal astrometry, the remaining refractive error is generally of the order of a few hundredths of a second of arc and has some systematic component. The use of multicolor observations, which is practical only in interferometric techniques, improves the situation, but not to the level of milliseconds of arc.

Atmospheric Turbulence. The atmosphere is not a smooth medium. Atmospheric stratifications move, and unstable vortices develop and evolve with time. They produce variations in refractive indexes and in the inclination of equally dense layers. The largest affect the angle of refraction. The dimension of the smallest turbulent cells are in the range of 5–30 cm and are produced by the temperature difference between the ground and the air and by the irregularities of the surface. They move with the wind so that the light is randomly deviated and seems to originate from different points in the sky. In addition, rays interfere, and the resulting instantaneous image of a star, called speckle, is deformed and moves rapidly around some central position on timescales of a few hundredths of a second. The resulting accumulated image is a disk whose size is at best $0.5''$ and is generally of the order of one second of arc on good nights, 2–3'' on others. These numbers characterize the visibility and indicate that, however small the theoretical resolving power of a telescope may be, the images are always larger than $0.5''$. So, whatever is the care with which the photocenter of such an image is determined, the pointing precision is necessarily limited to a few hundredths of a second of arc.

Mechanical Properties of the Instrument. The structure of a telescope, in particular of the transit, is subject to torques that depend on its inclination. In

practice, it is impossible to model it so that its effects introduce biases in determining refraction. In addition, again in the case of a transit, declination is determined by using a divided circle. The accuracy and the precision of the readout of the marks are limited, not to mention the deformations of the circle due to temperature. These effects bias the observations as a function of the time of observation during the night. These causes of errors, together with other perturbations specific to individual instruments, constitute an ensemble of limitations to the accuracy of astrometric observations that are of the same order of magnitude as those due to the atmosphere.

Sky Coverage. It is important for all global studies (kinematics and dynamics of the Galaxy, for instance) that positions and proper motions be referred to a single frame, independently of their situation on the sky. To achieve a global astrometric catalog, it is necessary to compile it from regional catalogs produced by many semiglobal instruments. Despite all of the efforts that are made to reduce the systematic differences among them, inevitably not all are corrected, especially if there are undiscovered correlations or similar systematic effects. The last—and best—global catalog is the FK5, produced in 1988 (12), that contains 1535 stars. The accuracies are about 1 mas in proper motions per year and $0.08''$ in position at the date of the catalog. The latter figure is an enhanced marker of the uncertainties of the proper motions used to update the positions from the mean epoch of observations (1950) to the present. The systematic differences with accurate space observations by Hipparcos, shown in Fig. 4, illustrate the complex structure of the biases due to the various causes described before. The figure also shows the actual intrinsic limitations of ground-based astrometry. Only astrometry from space can improve the situation significantly.

Space Astrometry. Going to space to perform astrometric observations has long been a dream of astrometrists. A principle that could be used was first published in 1966 by P. Lacroute (13). The method proposed was actually retained for Hipparcos and is now adopted for several future projects. However, at that time, space technology could not meet the accuracy challenges. Reproposed in 1973 to the European Space Agency (ESA), a feasibility study of such a mission was approved in 1976, and the project was included in the ESA mandatory science program in 1980. The project was delayed because of the priority given to the Halley comet space mission Giotto, so that the detailed design study was completed only in December 1983, and the hardware development started immediately after. Another delay was caused by a failure of the Ariane launcher, so that the actual launch of the satellite occurred in only August 1989.

What was required from space astrometry was first to eliminate the limitations described earlier. Clearly, the absence of atmosphere was the first objective, but at the same time, the possibility of homogeneously scanning the whole sky and giving rise for the first time to true global astrometry was achieved. In addition, the very small gravitational and radiative pressure torques that exist in space do not affect the shape of the instrument. Finally, in the absence of atmospheric turbulence, the shape of the star image is entirely defined by optics and can be modeled with extreme accuracy.

However, a new very serious difficulty appeared: how to monitor the orientation of the satellite. In ground-based astrometry, the orientation of the

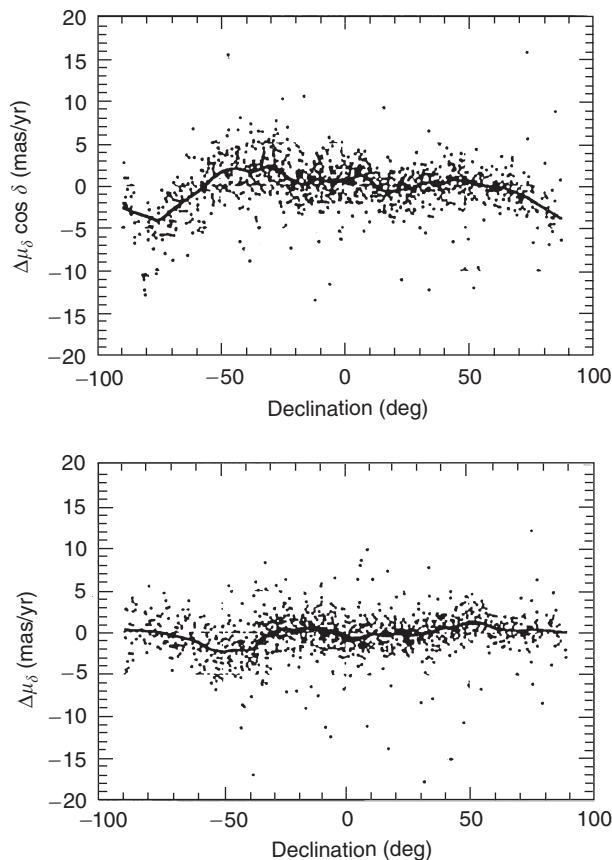


Figure 4. Differences in proper motions in right ascension and declination between the FK5 and the *Hipparcos Catalogue* as functions of declination. The solid line is a robust smoothing of the data.

instrument in space is a function of the parameters of Earth's rotation which are determined independently and very accurately by a specialized service, the International Earth Rotation Service (IERS). They include Universal Time from which sidereal time is computed, polar motion and, in space, precession and nutation. All are known with superabundant accuracy. In space, there is no such external reference and, at least for global astrometry, the orientation (or attitude) of the satellite must be determined simultaneously as accurately as the expected accuracy of star observations.

Until now (year 2000), two space astrometry missions were launched successfully. The *Hipparcos* mission was a global astrometric mission. The other comprises the astrometric facilities on board the Hubble Space Telescope (HST) all directed toward small-field astrometry. Both are described in the following sections. In addition, there are several space astrometric projects expected to be launched, if approved, during the first decade of the twenty-first century. Their principles are mentioned in the preceding section.

The Hipparcos Mission

Hipparcos is the acronym for high precision parallax collecting satellite, which points out the main astrophysical objective, but recalls also Hipparchus, the Greek astronomer who discovered precession and is the author of the first star catalog. The satellite was launched by ESA on 8 August 1989. A geostationary orbit was aimed at, but due to the failure of the apogee boost motor, the final orbit was very elongated; the perigee was at an altitude of 500 km, and the apogee was at 36,500 km. The period was 10 hours and 40 minutes. Communications with Earth were secured by three stations in Odenwald (Germany), Perth (Australia), and Goldstone (USA). They ensured direct visibility that covered 97% of the orbit and 93% of the useful observing time. Satellite control and pretreatment of the data were provided by the ESA Operation Center (ESOC in Darmstadt, Germany).

Because of the difficulties that arose from the change from the nominal to the actual orbit, the operations started only at the end of November 1989. They stopped in March 1993, after the failure of several onboard gyroscopes. Taking into account several interruptions, the total useful data collected represents an accumulated 37 months of observations. However, instead of quasi-continuous 24 hour a day observations anticipated for the nominal mission, only 7 to 9 hours and sometimes less observation time per orbital revolution could be achieved because occultation times by Earth, passage through radiation belts which induces strong Cerenkov radiations in the optics, and illumination by the Moon when it was near a field of view that produced a noise that masked the signal had to be excluded.

Principle of Hipparcos. The principle of Hipparcos is sketched in Fig. 5. The main characteristic is that two fields of view are focused on a single focal surface. The optical axes from the center of each field are combined by two glued half-mirrors called a beam combiner whose angle sets the angular reference γ , known

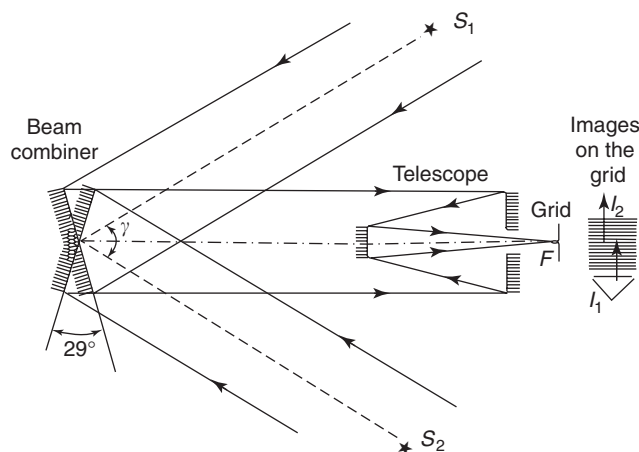


Figure 5. Principle of Hipparcos showing the motion of the images I_1 and I_2 of the stars S_1 and S_2 in different fields of view.

as the basic angle. On the focal surface is a grid composed of slits parallel to the intersection of the mirrors, a fundamental direction that we shall call vertical. The satellite revolves slowly around an axis parallel to this direction, and the light of the star is modulated by the grid. A photoelectric receiver registers the resulting signal. On both sides of the main grid are two systems of vertical and chevron slits, called star-mappers. Another photoelectric system registers the light crossing a star-mapper that provides data for determining the satellite's attitude. These data are also used for astrometry in the framework of the Tycho program. Note that, except in part for the star-mapper, the observations are one-dimensional, they amount to determining the transit time of star images through the grid. To cover the whole sky, it is necessary to modify the satellite's attitude in a predetermined way.

Description of the Hipparcos Payload. Construction of the payload followed the principles described before. The basic angle is $\gamma = 58^\circ 31.25''$. The beam combiner was produced by cutting a 29-cm Schmidt corrected mirror into two halves. The refractive Schmidt configuration was chosen to have a large astrometrically good field of view across more than 1.3° in diameter. The space structure of the dual telescope and its baffles that protect it from stray solar light is shown in Fig. 6. The light paths from the two fields to the optical block on which the grids are engraved are shown in the figure. The equivalent focal distance is 140 cm. A star produces a diffraction pattern on the grid that is elongated along the vertical direction. The along-track dimensions range between 0.5 and 0.7 seconds of arc in the sky, depending on the color of the star. The grids are engraved on the front side of the optical block which is curved so that it matches the focal surface of the telescope. The main grid consists of 2688 regularly spaced slits whose period is $8.20\ \mu\text{m}$ ($1.208''$ in the sky); the transparent width is $3.13\ \mu\text{m}$ ($0.46''$ in the sky). It covers two fields of view in the sky of $0.9^\circ \times 0.9^\circ$. The vertical extension of the star-mapper grids is limited to 0.7° . Each grid is composed of four $0.9''$ wide slits that have pseudorandom separations respectively of a , $3a$, and $2a$

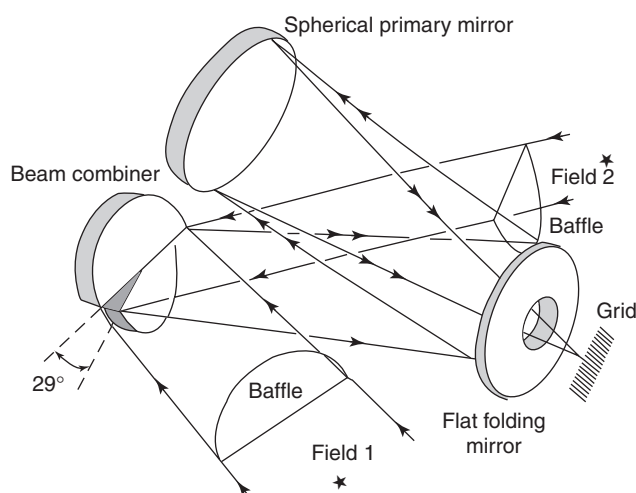


Figure 6. Space configuration of the Hipparcos optics.

($\alpha = 5.675''$ in the sky). Only the preceding star-mapper was operated. The second, which was redundant, was actually never used.

The receiving systems were quite different for the main grid and for the star-mapper. The rear side of the main grid was curved to serve as a field lens of the optical system used to image the full grid on the entrance of an image dissector. This special kind of photomultiplier produces an electronic image on the back wall of the tube. A small hole leaves the way open only to electrons coming from a tiny portion of the electronic image. A set of magnetic deflectors can be controlled to shift any point of the electronic image on the hole. As a consequence, only the light coming from a $30''$ radius circle in the whole focal image is recorded. The way this is used will be described later. In contrast, all of the light that enters the star-mapper grids is transmitted to a dichroic mirror that splits the light into two wavelength ranges that correspond roughly to the B and V filters of the Johnson UBV photometric system, called here B_T and V_T . Then each channel is directed toward a different photomultiplier. The photoelectrons are recorded at a rate of 1200 Hz for the main grid and 600 Hz for each of the star-mapper channels. More details on the Hipparcos instrumentation and operations are given in Volume 2 of the Hipparcos and Tycho Catalogue (14).

Scanning the Sky. Because observations are made along a scan in a narrow band 0.9° wide, it is necessary to modify the attitude of the satellite so that all of the sky is scanned as homogeneously as possible and all of the stars are observed for roughly the same amount of time. Various scanning laws can do this, but there are additional constraints. First, and this is the most important condition, the angle between the observed fields of view and the Sun must be at least 45° to minimize stray light. However, the inclination of the scan with respect to the ecliptic should be as small as possible. The parallactic deviation is parallel to the plane of the ecliptic. One wishes to maximize its projection along track because this quantity is used to determine parallaxes. So the inclination chosen is at the limit of acceptance by the first condition. Finally, the attitude should change slowly so that there are overlaps between successive scans.

Nominal Scanning Law. As a compromise between these conditions, the following nominal scanning law was adopted. The satellite rotates in 2 hours 8 minutes allowing 19 seconds for each star to cross the main grid. The rotational axis circles the direction of the Sun in 57 days, keeping an angular distance of 43° from the Sun. Figure 7 shows the motion of the axis of rotation in one year and the part of the sky scanned in 70 days.

Left to itself, the rotational axis would drift rapidly due to the various torques that are applied to the satellite (gravitation, radiative pressure, reaction of the gyroscopes, etc.). To follow the scanning law demands active attitude control that is realized by six gas-jet thrusters using compressed nitrogen. The satellite attitude is monitored onboard (see earlier section). When it deviates by 10 minutes of arc from the nominal scanning, gasjets were actuated to reverse the natural attitude drift. In practice, this happened four to six times an hour during the observation conditions. When the satellite was in Earth's shadow or in radiation belts, attitude control was sometimes very bad, special scanning law recovery procedures had to be applied, and observations were not possible during these maneuvers.

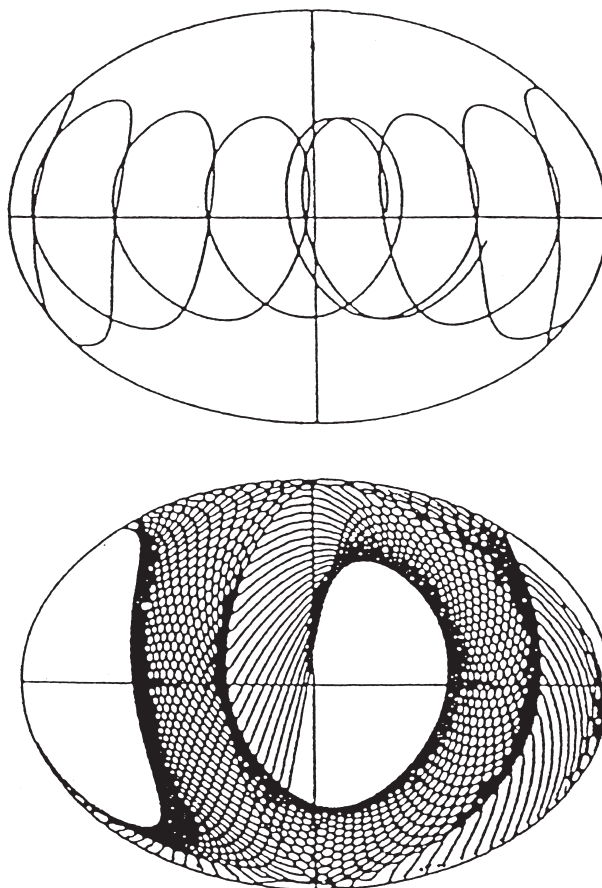


Figure 7. Hipparcos sky scanning law in ecliptic coordinates. *Above:* Motion of the satellite axis in one year. *Below:* Part of the sky scanned in 70 consecutive days.

Onboard Attitude Determination. The attitude had to be known continuously during the observations at an accuracy better than one second of arc. The onboard attitude determination used five rate-integrating gyroscopes calibrated in real time using star-mapper data from star crossings. When, a star image crosses a star-mapper slit at a time t , one may write an equation stating that the star lies on the projection of the slit in the sky. Knowing the celestial coordinates of the star corrected for stellar and orbital aberrations, the equation sets one condition on the parameters describing the attitude at time t . If the evolution of the attitude between two gas-jet actuations is smooth, a few equations of condition are sufficient to calibrate the gyroscope drifts and therefore to know the attitude with the desired precision.

The Input Catalogue. As just stated, the necessary real-time orbit was determined on the basis of rough knowledge of the position of some stars. Actually, there was also the same need in the operations of the main grid: to shift the image of a star on the hole of the image dissector, it is necessary to know the position of the star as well as the attitude of the satellite. In addition, because the

main grid is periodic, there is no other method to know on what slit the image of a star lies than to know the position significantly better than the grid period. Finally, if all of the observable (up to magnitude 12.5) stars were to be observed, the time allocated to each star would have been insufficient. This led to a limit of the number of stars in the observing program and to creating a list of program stars. Finally, 118,322 stars were selected after a considerable amount of optimization between the astrophysical return and the operational requirements of an even distribution throughout the sky.

From this list of stars, an *Input Catalogue* was constructed by an international consortium led by C. Turon (15). This catalogue provided the positions of the program stars to a mean uncertainty of $\pm 0.25''$. The preparation involved a large number of astrometrists, most of whose work had to be reobserved on transits or measured on photographic plates. To optimize the observation time, the magnitudes of the stars also had to be known to ± 0.5 magnitude. This objective involved many photometrists, particularly because the time variations of irregular, long-period large-amplitude, variable stars had to be monitored, work that was pursued during the entire mission. Finally, the Input Catalogue provided other known parameters of the stars such as the description of double and multiple stars, many of which were reobserved on this occasion, variability types and amplitudes, parallaxes, proper motions and various identifications in major star catalogues. If Hipparcos had not been successful, the Input Catalogue would still have been a very useful database for many astronomical investigations.

The Hipparcos Data Reduction

Reduction of the Hipparcos data was undertaken by two different international consortia (FAST led by J. Kovalevsky, and NDAC led by E. Høg and later by L. Lindegren). It is a very complex process that cannot be presented in the scope of this Encyclopedia. Details of the methods used by each consortium can be found in several publications: Reference 16 for FAST, Reference 17 for NDAC as well as a review on Hipparcos by Van Leeuwen (18), and a very detailed account of both is given in Volume 3 of the final catalog (14). So, we present here only a sketch without giving the rationale and the mathematics of the methods adopted.

Roughly, the reduction procedure followed by both consortia is divided into three steps. In the first, only the data acquired during one orbit were considered. During that time, the scanning law restricts the observations to a band in the sky that had a maximum width of 3° . All of the observations were treated together and projected on a single celestial great circle, called the Reference Great Circle (RGC), chosen in the observation band. The abscissas of all observed stars were reckoned from a conventional origin on the RGC. In practice, this step included analyzing photon counts on the main grid and determining the positions on the grid at some defined times, analyzing photon counts on the star-mapper and calculating the attitude, and finally computing the abscissas on the RGC. Some details of these three procedures are given in the following subsections. The second step involves a large number of RGCs to shift the origins of individual RGCs, so that they form a single consistent reference system. In the third step,

the RGC abscissas of a given star are combined to determine the five astrometric parameters, namely, the position at a common reference epoch (1991.25), the parallax, and the two components of the proper motion. Later, the three steps are iterated to improve the solution. In addition, several off-line tasks were performed to determine the elements of double and multiple stars and the magnitudes and positions of observed minor planets. Finally, the results obtained by the two consortia were merged and rotated to match the International Celestial Reference System (ICRS).

Reduction on a Reference Great Circle. As stated before, Hipparcos data reduction consists of three distinct phases that are briefly presented here.

Grid Coordinates. They are computed for the central time of a 2.133-second frame during which main grid photon counts are registered. Each count is the integral of the modulated light curve during 1/1200 second. The counts are used to compute the intensity modulation in the following form:

$$I = I_0 + B + I_0 M_1 \cos(\omega t + \phi_1) + I_0 M_2 \cos 2(\omega t + \phi_2), \quad (13)$$

where I_0 is the mean intensity of the star, B is the background noise, M_1 and M_2 are the modulation coefficients, and ϕ_1 and ϕ_2 are the phases reduced to the mean time of the frame. For a single star, $\phi_1 = \phi_2$, but this is no longer the case for double stars for which the observed intensity is the sum of the intensities of the components. Figure 8 shows some shapes of modulation curves. A conventional phase,

$$\phi = C_1 \phi_1 + C_2 \phi_2, \quad (14)$$

with

$$C_1 + C_2 = 1, \quad (15)$$

was used to represent the position of the star within a grid step. Then, the horizontal coordinate on the grid is

$$X = Ns + s\phi/2\pi, \quad (16)$$

where $s = 1.208''$ is the grid step. The integer N is deduced from what is known about the coordinates of the star and the attitude of the satellite. It may be in error by ± 1 , rarely more. This is called a grid-step error. The vertical Y coordinate is computed from the same data and a field-to-grid transformation that is calibrated separately.

Attitude Determination. The principle of the attitude determination has already been given earlier. During the reduction phase, it is almost uniquely based on star-mapper observations. See References 19 and 20. The photon counts recorded during the transit of the star image through one of the grid systems are correlated with a calibrated response curve obtained by analyzing many transits of single stars. This gives the transit time through a conventional mean line for which the condition equation mentioned earlier is written. For the interval of time between two gas-jet actuations, the attitude varies smoothly and can be represented by some analytical formula (trigonometric series, polynomials, or

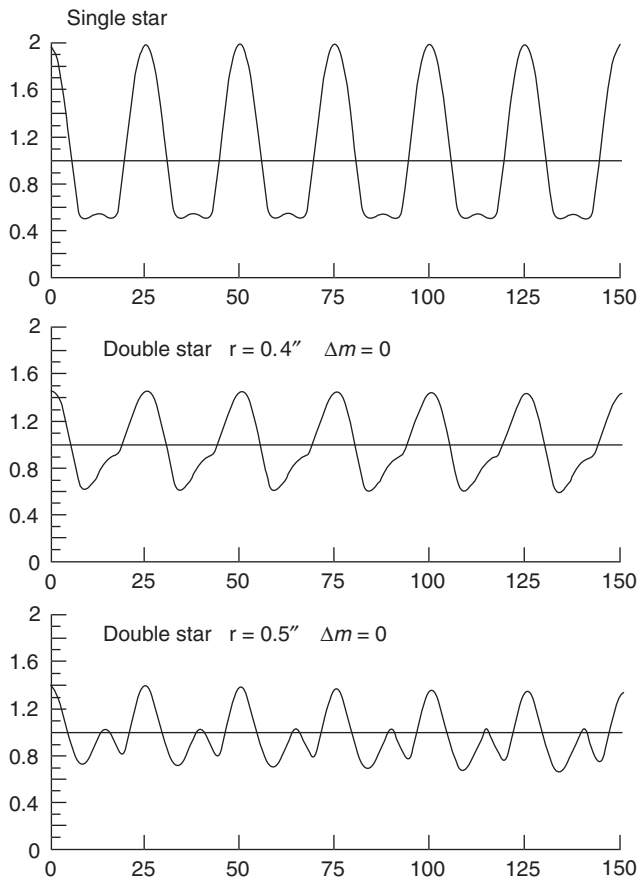


Figure 8. Theoretical modulation curves of a single star and two different configurations of double stars that have equal luminosity components.

cubic splines). The parameters of these formulas are determined from all of the equations of condition in the interval, and the result is used to interpolate the attitude for any time.

Abscissas on the RGC. This is a rather long and complex task (21). First, the celestial coordinates of the star are deduced from the grid coordinates and the attitude, using grid-to-field transformations, which are third-order polynomials linking the X and Y coordinates to the tangential coordinates in the sky at the central point of each field of view. The parameters A_i of the transformation also depend on the color and the magnitude of the star. Then, the position so determined is projected on the RGC giving an abscissa which therefore depends on the value of the attitude at the time of the grid coordinate, the parameters of the grid-to-field transformation, the coordinates of the star, and the basic angle γ . In the equation that represents this dependence, the attitude along the RGC is considered unknown and is represented by the coefficients of an ensemble of cubic splines that cover the entire set of observations. Some 1500 parameters are needed to cover it. Errors in other components of the attitude as

well as the vertical coordinate of the star have only a limited influence on the error budget. So they are considered known and will be improved only during the iterations.

One abscissa observation provides one equation: 1000 to 1500 stars are present on a RGC. Each star is observed several times during an orbit, and generally nine times during one grid transit. Overall, this represents of the order of 40,000 equations that have about 3000 unknowns, including the field-to-grid transformation coefficients and a correction of the basic angle. The system of equations is largely overdetermined, and all of the unknowns are straightforwardly determined by the least squares method. The values of the coefficients A_i are kept as calibrated values of the grid-to-field transformation and its inverse, the field-to-grid transformation. The fact that the basic angle is determined (actually with an uncertainty of 0.2 to 0.3 mas) shows that, in reality, the yardstick for angles is not this material artifact, but 2π in the attitude determination. However, its existence is fundamental because it is the basis of the rigidity of the solution. After iterations, the uncertainty of the abscissas ranges between 3 mas for bright stars to 5 mas for fainter stars. The along-track attitude is obtained with a similar uncertainty of a few mas.

Sphere Solution. Once a sufficient number of RGCs is processed and covers the whole sky, the second reduction step can be undertaken. Each star, observed at different times, moves in the sky, as shown in Fig. 9. The motion is the combination of parallactic displacement and proper motion. Therefore, the abscissa

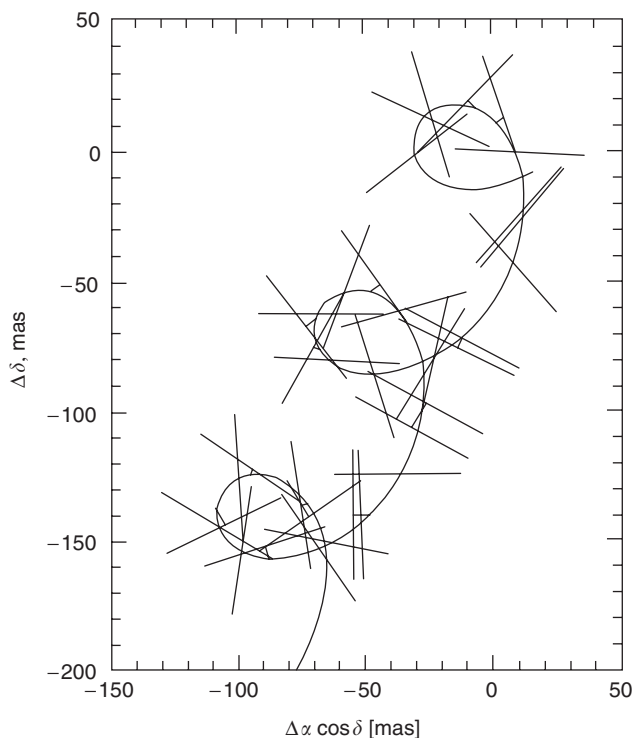


Figure 9. Path of a star and the projections on the RGCs on which it was observed.

of the star on an RGC is a function of five astrometric parameters and of the inconsistency of the position of the origin of the RGC. If one considers not the position, but a correction to the position in the reference catalog, one can write a linear equation involving the five star parameters and a correction $\Delta\alpha_j$ to the position of the origin of the RGC(j). At this stage, it is not necessary to take all of the stars, so only the 40,000 well-observed single stars are considered. At the end of the mission, each star was observed in the mean between 20 and 40 times, so that there is a total of more than a million equations with 200,000 star unknowns and 2500 $\Delta\alpha_j$. If the star unknowns are eliminated, a large linear system that has 2500 unknowns remains which is solved by the method of least squares. With the new origins, the RGCs constitute a rigid mesh that represents the provisional Hipparcos reference system.

Determination of the Astrometric Parameters. The same equations are now written with the new RGC origins for all stars and sorted star by star. There are generally 20 to 40 equations with five unknowns that are readily solved, and the astrometric parameters are obtained with the complete variance-covariance matrix. In some cases, the residuals of the solution show that the proper motion linear model is not adequate. A polynomial representation is then adopted, meaning that there is an unseen companion that perturbs the motion of the star. It is thus classified as an astrometric double star. Generally, the companion is a faint star, but in a few cases, it was shown that it was a brown dwarf (22).

If the star is recognized as double or multiple by examining its modulation curve, the parameters of the system (magnitudes of each component and their relative positions) must be determined before, or together with, the computation of the astrometric parameters. In the great majority of cases, the assumption that it is a double star is found true, and the magnitudes and the relative position are determined from the time variation of the modulation coefficients. Then, the equations for astrometric parameters are rewritten and solved with an especially adapted algorithm.

Final Steps and Results. The ground-based data used to initialize the reduction process, to determine the attitude, for the first time and to compute the grid coordinates are too inaccurate to be kept throughout the reduction. They must be improved to suppress the effect of their errors on the final result. This is done by iterating the reduction using the latest calibrations and star astrometric parameters obtained in the preceding reduction. Actually, the first complete solution could be obtained with 18 months of data, and this already greatly improved the Input Catalogue information. By performing an iteration, the treatment of photon counts did not need to be redone, so in the first step only the grid coordinates, the attitude, and the reduction on the RGC had to be repeated. The last two steps were of course totally reprocessed.

Merging the Catalogs. When both consortia had finalized their last catalog, the question arose how to present a single Hipparcos set of results to the scientific community. The idea of just making a weighted average of the astrometric parameters obtained by the two consortia was rejected because there was no statistically justified means to compute the resulting correlations between the merged parameters. Although the consortia were justified in considering the observations as independent, the two solutions obtained were highly correlated. So, it was decided to return to step 2, take the abscissas obtained by the consortia

together, and treat them as correlated observations (20). The correlations were computed for a number of stars, and an analytical representation of these correlations was determined as a function of time, magnitude, and the standard errors of abscissas obtained by each consortium. Applying this formula, steps 2 and 3 were reprocessed taking these correlations into account. The standard errors and correlation coefficients obtained this way are those published in the Hipparcos Catalogue.

Link to the ICRS. The system to which the merged positions and proper motions were referred was close to the FK5 system. The objective was that the reference system should be the ICRS. To do this, it was necessary to determine a rotation of the Hipparcos reference that would fit it to the ICRS or, better, its realization, the ICRF. Actually, two rotations are necessary: one for the reference epoch and another proportional to time.

Many astronomers contributed to this task by providing positions and/or proper motions of Hipparcos stars with respect to extragalactic objects. Several ground-based techniques were used:

- observations of radio-stars by long-base or connected radio interferometry with respect to quasars of the ICRF,
- photographic plates taken at different epochs providing proper motions referred to galaxies, and
- data taken from several catalogs of proper motions referred to galaxies.

In addition, direct observations were made at the astrometric focus of the Hubble Space Telescope. A general weighted solution was made to determine the rotations (23) which were then applied to the merged catalog. The resulting astrometric parameters are those published in the Hipparcos Catalogue. The uncertainty of the rotations that provide this link to the ICRS is 0.6 mas for the position at epoch and 0.25 mas/yr for the time-dependent rotation.

Accuracy of the Hipparcos Astrometry. The published catalog (14) contains astrometric parameters of 117,955 stars. The median uncertainties for stars brighter than magnitude 9 (two thirds of the stars) are the following:

- ± 0.77 mas in right ascensions ($\Delta\alpha \cos \delta$),
- ± 0.64 mas in declinations,
- ± 0.97 mas in parallaxes,
- ± 0.88 mas/yr in proper motions in right ascension ($\mu_\alpha \cos \delta$),
- ± 0.74 mas/yr in proper motions in declination.

Several attempts were made to evaluate possible biases in the results. Several tests were made, including the parallaxes of stars in the Magellanic Clouds, O- and B-stars in clusters, and a model of distribution of negative parallaxes (24). All concluded that a general bias in parallaxes should not exceed 0.1 mas. However, very locally, and particularly in clusters where the observations are partially correlated, larger systematic errors may exist, but in any case no more than a few tenths of a mas.

Another set of results concerns double and multiple stars. Using the modulation curves, one can detect them and solve for the respective positions and luminosities of the components of double (25) and sometimes even multiple stars (26). As a result, 12,430 double or multiple systems were solved, out of which almost 3000 are newly discovered systems. One must also mention the discovery or confirmation of 2910 astrometric double stars, although 8542 stars were suspected to be nonsingle, but no solution was found.

The results of the Hipparcos mission were immediately used in a large number of investigations in all fields of astronomy and astrophysics. The first results are published in Reference 27. A synthetic account is given in Reference 28.

Hipparcos Photometry. Although nominally Hipparcos was an astrometric mission, the payload was a remarkable photometer. The analysis of modulation coefficients by different methods (29,30) also provides the magnitudes of the stars. The system of Hipparcos magnitudes is a wide-band photometric system defined by the transmission of the optics and of the image dissector. A great effort in calibration had to be made to correct for the inhomogeneities of the sensitive surface and aging of the optics and the detector. The sensitivity of the latter decreased with a marked chromatic dependence. One also had to estimate and remove the effect of the background. The calibrations were based upon about 22,000 standard single stars all of which had multicolor photometric observations in various systems. The resulting Hipparcos magnitudes were derived by M. Grenon in Geneva.

The photometric data were reduced by the same two consortia. After numerous comparisons, the merger of the results simply consisted of computing the mean. The results are published in the same catalog as the astrometric results. There are 118,204 entries. The median uncertainty for stars of magnitude smaller than 9 is ± 0.0015 magnitude. In addition, the Hipparcos Epoch Photometry Annexes, available in machine-readable form, give 13 million individual measurements, one per main grid transit. A total of 11,597 stars has been recognized as variable, out of which 8237 have been discovered as such from Hipparcos data. Among them, 2712 were recognized as periodic and for most of them, the catalog provides folded light curves in addition to light curves for 1101 other objects.

The Tycho Project. Only a very small part of the observations made with the star-mapper are used in the Hipparcos data reduction for determining attitude. Actually, all of the stars that appeared in both fields of view gave signals that were recorded. The objective of the Tycho project was to recover all of these data and use them to obtain the astrometric and photometric information they contain. The data were treated by the international consortium TDAC led by E. Høg. An overview of the reduction method is given in Reference 31, and a detailed description is given in the Hipparcos and Tycho Catalogue, Volume 4 (14).

Principle of Tycho. We have seen earlier that when a star image crosses the mean line of one of the grid systems at some time t , one may write an equation that links the position of the star in the sky and the attitude. If the position of the star is known, the equation constrains the attitude at time t . Conversely, if the attitude is known, the same equation becomes a constraint to the position of the star. Assuming that the position of the star is known

approximately, the condition is represented on the plane tangent to the sky by a straight line parallel to the slit. Each grid crossing produces one such line. A shift perpendicular to the line corresponds to a difference in transit time. Figure 10 shows how these lines roughly converge near a point that represents the actual position of the star. A few lines correspond to misidentified transits. The problem is to find this point from these lines. Before, one had to identify the lines that pertain to the same star. This identification is actually the most difficult part of the reduction.

Prediction and Identification of Transits. As in the case of Hipparcos, but for different reasons, it was necessary to have a Tycho Input Catalogue (TIC). It first contained the 3.26. million brightest stars in the sky to a limit of Johnson $B=12.8$ or $V=12.1$ magnitudes and was constructed by merging the Hubble Space Telescope Guide Star Catalog (see later) and the Hipparcos Input Catalogue.

The transit times obtained during the first months of the mission were compared to the predictions for all stars of the TIC computed using the first versions of the attitude obtained by the Hipparcos data reduction. This showed that more than 60% of the TIC stars were not found probably because most of them did not satisfy the thresholds for acceptance of a transit. The remaining stars constituted a new revised TIC (TICR) that contains some 1.26 million stars. Then the transit times for all of the objects of the TICR were predicted using the best available Hipparcos attitude description provided by the Hipparcos consortia and applying, of course, a grid-to-field calibrated transformation and the corrections for stellar and orbital aberrations.

The identification of transits consisted of pairing the observed transit time with that predicted (32). The accuracy of the predicted positions was around 0.2 second of arc. When a transit differed by more than $1''$ from the prediction, the

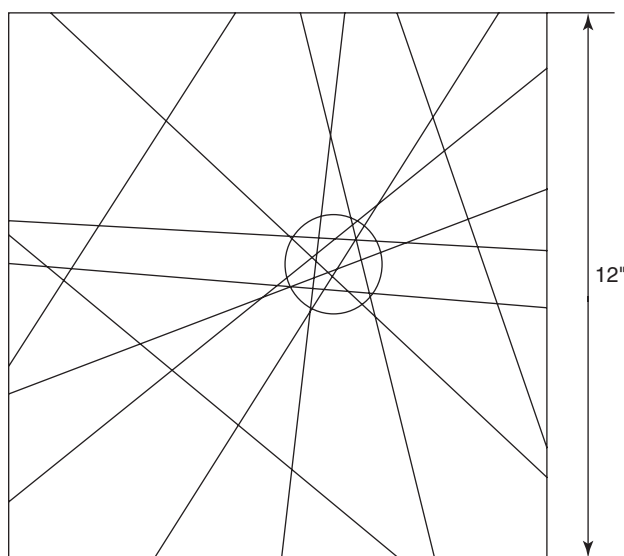


Figure 10. Loci of positions of a star derived from Tycho observations. Three lines correspond to misidentifications.

transit was rejected and later used for a search of companion stars. After processing these data, 6600 new companion stars were discovered.

Astrometric Reduction. The relation between the transit time and the position of the star is a transformation that involves, for each type of grid (vertical, upper, and lower chevrons) and each field of view, a set of coefficients that describes the grid-to-field transformation and corrections to the astrometric parameters given in TICC (33). The grid-to-field transformation was periodically calibrated using the basic equations applied to observations of Hipparcos stars whose positions were known with a superabundant precision for this particular objective. It included a global transformation analogous to that used for the main grid plus 68 medium-scale values as functions of the positions on the grids.

The determination of the astrometric parameters again used the same equations, but with the calibrated values of the field-to-grid transformation at the time of the observation. All of the identified transits of the star were collected and the position of the star described by its five astrometric parameters, as for Hipparcos, was determined by minimizing the sum of the squares of the distances to the lines shown in Fig. 10.

The final Tycho catalogue (14) contains 1,058,332 entries and is practically complete up to V_T magnitude 10.5. The median astrometric precision is 25 mas, if one considers all stars, and 7 mas if only bright stars ($V_T > 9$) are considered. Evaluation of systematic errors showed that they are smaller than 1 mas. This is, of course, far from the precision of the Hipparcos Catalogue, but the gain is that it concerns nine times more stars. Actually, the difference in precision is easily explained if one considers that the cumulated observation time for a star is 25 to 30 times smaller than that on the main grid. It is expected that a new reduction of the data now in progress with lower thresholds for the acceptance of transits may add another million stars of magnitudes V_T in the range of 10.5 to 11.5.

Photometry of Tycho. The data collected by each of the two channels are treated separately in a manner similar to the Hipparcos photometric observations (34). The calibrations were performed from the observations of about 10,000 standard stars with magnitudes between 4.5 and 9 and included dependences on position along the slits, field of view, star color, and channel. The median photometric precisions are as follows in magnitudes:

- For all stars:
 ± 0.07 in B_T , ± 0.06 in V_T , and ± 0.10 in color index $B_T - V_T$.
- For bright stars ($V_T < 9$):
 ± 0.014 in B_T , ± 0.012 in V_T , and ± 0.019 in color index $B_T - V_T$.

Astrometry with the HST

The Hubble Space telescope (HST) is not primarily designed to perform astrometric observations, but to support several scientific instruments whose common requirement is that the telescope must be able to point in any given direction in the sky and stay pointed with very high stability as long as necessary (35). This task is allocated to three fine guidance sensors (FGS). Actually two FGSs are

sufficient to locate a target and stay pointed at it. The third one remains free with the possibility to do astrometry within the field of view. So, in general, astrometric measurements are confined to a certain field in the vicinity of the region studied by other instruments. However, some astrometric programs are scheduled for their own sake, and then the choice of targets is left to the discretion of the astrometrists. Occasionally, one of the scientific instruments, the wide-field planetary camera (WFPC) is used to perform, despite its name, very narrow field astrometry. In even more exceptional cases, the faint object camera (FOC) may also be used for this purpose.

Figure 11 shows how the focal surface of the HST is divided among the various instruments. The WFPC occupies the central part, surrounded by four fields from which the light is transferred to four scientific instruments. The three FGSs occupy the outer ring of the focal surface. The fields allocated to the FGSs are three 90° segments of an annulus of inner and outer radii that correspond to 10.2 and 14 minutes of arc in the sky.

It is well known that the main mirror has an important spherical aberration and that the secondary mirror is slightly tilted and decentered. In addition, there was an important jitter due to the excitation of the solar panel assemblies when the satellite passed into or out of direct sunlight. All of this significantly impaired the astrometric quality of the telescope. The Hubble repair mission in December 1993 suppressed the jitter and the WFPC was replaced by a new camera with modified optics to correct the defects of the telescope. The faint object camera, was corrected by the additional optics provided by the

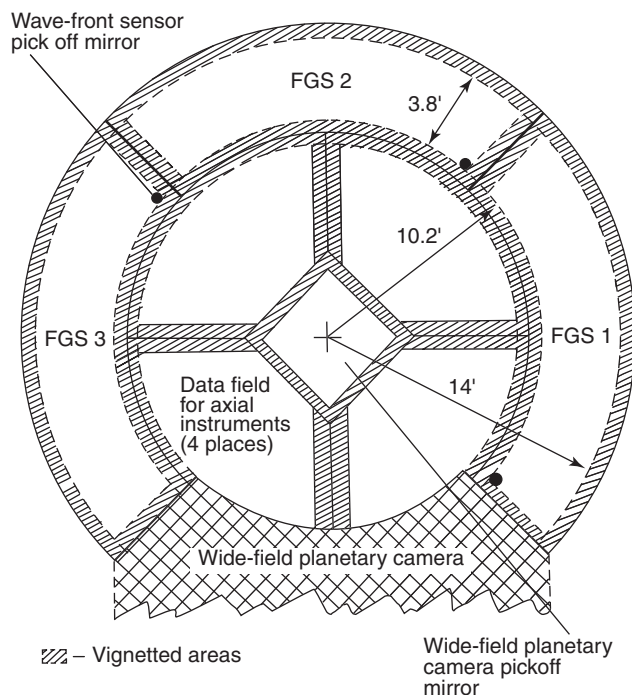


Figure 11. The focal surface of the Hubble Space Telescope.

multicorrector COSTAR. But the FGSs remained untouched, and the only improvement came from replacing the solar cell arrays that suppressed the jitter. Otherwise, the situation remained characterized by the major spherical aberration added to the expected astigmatism inherent in a Ritchey–Chrétien telescope configuration (a configuration chosen to avoid an important coma that would have been a more severe penalty to astrometry from the FGS). The point-spread function has significant features even a few seconds of arc away from the center of an image and in addition depends strongly on the position in the field of view. Only 15% of the light is concentrated in the $0.1''$ central circle instead of more than 50%, as anticipated (36). This corresponds to a loss of more than one magnitude in access to fainter stars and some general degradation of astrometric capabilities. However, even it could have been better, the FGS with careful calibrations, remains a remarkable tool for astrometry.

Description of the FGS. The light that reaches the focal surface is deflected by a pickoff mirror and an aspheric collimator into a first star selector that provides exact collimation and correction for nominal astigmatism, spherical aberration, and field curvature. When its two mirrors rotate about an encoding axis, it produces a rotational angle θ_a at a fixed calibrated deviation angle δa that leads to point T (Fig. 12). A second selector produces a rotational angle θ_b around axis T at another calibrated deviation angle δb . The composition of these two rotations allows reaching any point of the FGS and selecting around it a $5'' \times 5''$ instantaneous field of view in which the fine pointing is performed. The coordinates of this point are

$$x = \delta a \cos \theta_a + \delta b \cos \theta_b, \quad (17)$$

$$y = \delta a \sin \theta_a + \delta b \sin \theta_b. \quad (18)$$

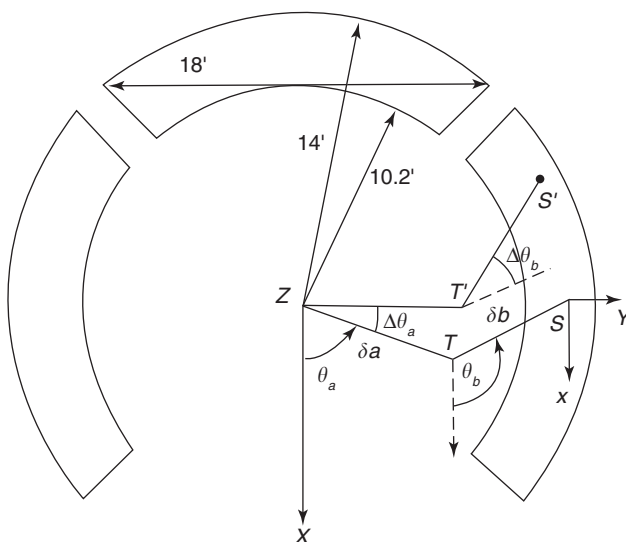


Figure 12. Coordinate systems of the FGS and relative measurements.

The image of this field is divided by a beam splitter and sent to two interferometric Koester prisms which sense, in two perpendicular directions, whether the incoming wave front is parallel to the front side of the prism. The outputs of each prism are sensed by photomultipliers. If I_1 and I_2 are the intensities sensed, an error signal S is produced:

$$S = (I_1 - I_2)/(I_1 + I_2). \quad (19)$$

It is null if the wave front is exactly parallel to the entry face of the prism. If both signals are null, pointing is accomplished; otherwise, a signal is sent to the selectors to correct the pointing. The value of S as a function of the depointing angle is the transfer function. Figure 13 shows the shape of the transfer function for images of increasing diameters. It pictures the effect of enlarging the image as the point-spread function widens. Because of the dependence of position on the optical properties, the transfer function must be calibrated by scanning across a number of known single and double stars of different colors in different spots of the field.

HST Pointing. Two FGSs are dedicated to point the telescope to guide stars. Pointing to two stars is sufficient to ensure a unique direction of the axis of the telescope. Because of the reduced fields of view, it was necessary to have a very dense ensemble of star positions all over the sky. This was the objective of preparing the Guide Star Catalog (GSC).

The Guide Star Catalog. A list of star positions and magnitudes of some 20 million stars in the 9 to 15 magnitude range (37). It is based upon micro-densitometer scans of Schmidt telescope photographic plates taken by the 48-inch Palomar Schmidt telescope and the UK Schmidt telescope in Australia. These plates were measured with a precision of $0.25''$, but the positions of reference stars for the reduction of the plates were taken from an old catalog (SAO Catalog published in 1966 and hence compiled from much older observations).

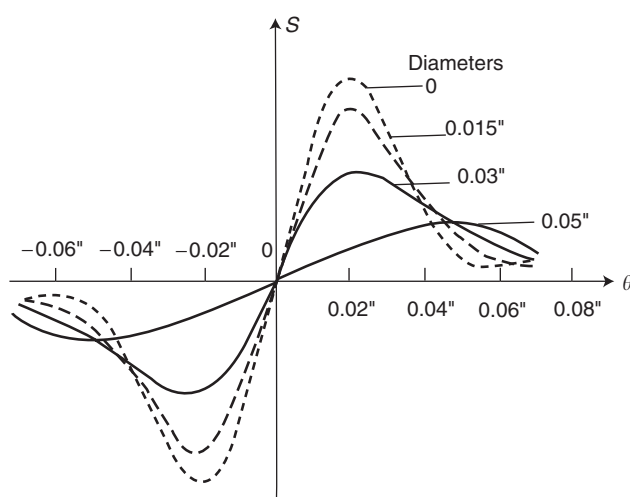


Figure 13. FGS transfer functions as functions of the diameter of the image.

The proper motion uncertainties were such that the positions in 1985 were strongly degraded, and the GSC had a mean uncertainty of 1.5 to 2 seconds of arc. However, this was sufficient for HST pointing.

HST Pointing Procedure (38). In the first stage, the telescope is roughly pointed toward the target by its spinning reaction wheels, using the references provided by the gyroscopes. Three fixed-head, star-trackers, independent of the telescope, are used to determine the position and the corresponding FGS coordinates of the two candidate guiding stars. The attitude control system keeps the direction of the telescope fixed, and a first FGS starts searching the guide star using a spiral scan from the indicated coordinates. When it finds it, the second FGS searches the second star. When it also finds it and if the relative positions are compatible with the GSC, the acquisition is confirmed. Then, the Koester prism interferometers send error signals to the attitude control system so that the error signals remain null. The stability of the pointing is about 7 mas rms, limited essentially by the reaction wheel jitter. Because of the uncertainty of the GSC positions, and the limited precision of the head star-trackers, the target may not be exactly in the center of the field of view of the active instrument. This is unimportant for the WFPC and the FOC but may be too large for other focal instruments. Therefore, all have a method of automatically centering the object or may send a signal to the ground-based control station which sends correction instructions.

Astrometry with the FGS. The third FGS does not differ from the two used for pointing, except for a different choice of filters. Once the telescope is rigidly pointed, as described in the preceding section, astrometric measurements may start but are restricted to the field of view of the third FGS, which is 3.8' wide and the maximum distance between objects cannot exceed 18'. The preparation of an astrometric observation is very complicated. First, it is necessary to determine how to place the field of view to include all objects of interest. From this, the direction of the optical axis of the telescope is to be determined, the guide stars chosen for each FGS, and their positions in the fields of view are computed in terms of the rotational angles θ_a and θ_b . Similarly, the positions of the targets must be defined in the same manner, so that the instantaneous fields of view of the Koester prisms are automatically set in the correct place. Finally, the order in which the objects are to be observed must be given together with the duration of each observation. There are three modes of observation (39).

The Lock-On Mode. In this mode, objects are successively viewed by interferometers, and the x and y coordinates are determined for each of them when the corresponding error signals are zero. Each star is measured several times in different sequences to avoid possible systematic effects. The positions on the FGS have then to be transferred onto the sky by a transformation that must be calibrated. This includes several successive calibrations.

- The distortion of the optical field angle is obtained by measuring 25 to 30 stars in two crowded fields (star clusters). This transformation is described by a polynomial.
- The plate scale calibration is obtained by measuring the positions of an asteroid as it moves through the field of view with respect to the background

stars which are also measured. The motion of an asteroid in a short time is very well known and serves as a standard of angle in the sky.

- The filter wedge error occurs when objects of different magnitudes are to be measured one with respect to another. The light of the brightest one is dimmed by a filter which may be slightly inclined. This is calibrated like the distortion of the optical field angle on a bright star cluster observed through different filters.
- The lateral color effect is produced by misalignments of the optics and a few mas chromatic effect has been observed. This calibration is accomplished by observing a couple consisting of one blue and one red star in different directions and positions in the field.

After applying the results of these calibrations, one must, in addition, correct for differential orbital aberration because the observations are not simultaneous and the orbital motion and hence the space velocity of the space vehicle change. Only then can one deduce the differences in right ascension and declination between the objects. The final uncertainties obtained are of the order of 3 mas for bright stars (magnitudes 0 to 15) and then degrade rapidly till the observing limit of magnitude 17.

The Transfer Function Mode. This mode is aimed at obtaining information about the structure of an object (40). Figure 13 has shown that the transfer function depends on the apparent diameter of the object. Similarly, it also depends on the shape of the object and particularly whether it is a double star. Calibration of the transfer function in various conditions was used to check the models describing it for different geometries of double stars.

In this mode, the object is slowly moved diagonally across the instantaneous field of view, and the error function is recorded as a function of the target's position. The analysis of the transfer function thus obtained with the help of calibrated models provides the relative positions and luminosities of the components. One may similarly obtain the diameter of the star, but note that the results obtained by the HST in the transfer mode are not better than those obtained from the ground by speckle interferometry and are less precise than those using Michelson interferometry. However, the great advantage of the HST is that they concern much fainter stars.

Moving Target Mode. This mode is used to track moving objects like minor planets. The FGS keeps locked on the object, and its position is periodically sampled. The preparation of observations in this mode must describe the expected path through the field of view.

Wide-Field Planetary Camera. The field of view used by the WFPC is a $3' \times 3'$ field centered at the optical axis of the telescope and then deflected to the relay optics and the receiving units. The original WFPC observation capacities were particularly hampered by telescope aberrations, so that it was changed during the repair mission. The new one is corrected for them and now has its nominal performance (Fig. 14).

The incoming $f/24$ light beam passes through a filter and a shutter and is then focused on a shallow four-faced, mirrored pyramid that can be locked into two positions (41). In one of them, it splits the field of view into four quadrants,

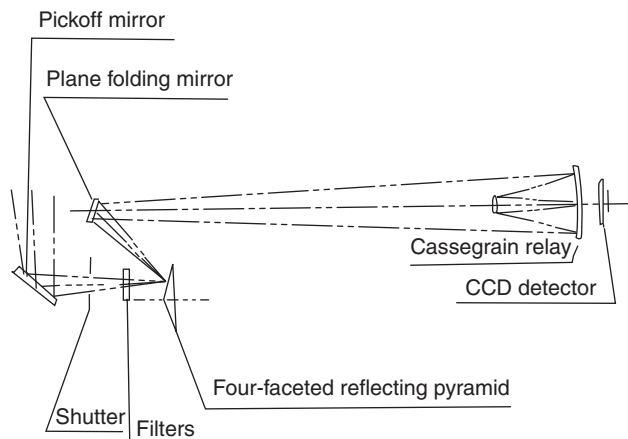


Figure 14. Optical path in the wide-field planetary camera.

and each is directed toward Cassegrain telescopes that convert them into $f/12.8$ beams focused on four 800×800 pixel CCD arrays that provide a total field of $160'' \times 160''$ in the sky. This configuration is the wide-field camera. To tie the four images collected into a single image, the pyramid has some uncoated small spots along the ridges. They are backlit, and the light coming through these spots produces markings on adjoining CCDs, providing fiducial points for the connection. Another $1.23''$ diameter nonreflecting Baum spot on the pyramid attenuates the luminosity of the object falling into it by about six magnitudes, to reduce the stray light if it is a bright star that allows surrounding faint objects to appear.

If the pyramid is rotated by 45° , the original beam is directed into four other Cassegrain telescopes that convert them into $f/30$ beams focused on four other 800×800 pixel CCDs. The field of view is reduced to $68'' \times 68''$ in the sky. This is the planetary camera configuration.

The observing sequence is quasi-identical to that followed in ground-based telescopes equipped with CCDs. A preflash is executed to wipe off ghost images that may have remained from the preceding exposure. To reduce the dark current, the CCDs are cooled below -95°C by thermoelectric coolers. Several preliminary calibrations must be made. One is the field-to-array transformation described by a third-order polynomial of the coordinates for each filter and the respective position. This is done by observing cluster stars previously measured with the FGS. Other calibrations include drawing a sensitivity map of the arrays and of the thermal noise. This is done on Earth but is checked in flight.

The reduction of the data allows obtaining positions to about one-fortieth of the pixel size, that is, about 5 mas for the wide-field camera and 3 mas for the planetary camera. This is comparable to the accuracy of the FGS, but the use of the WFPC is different; although the field of view is much smaller, the limiting magnitude is 21 for a few second exposure and fainter for longer. The astrometric objectives concern precise proper motions for recognizing astrometric binaries. Simultaneous observations of the same field by the WF/CA and the FGS are not possible because they see different parts of the telescope field of view, but observations of the part of the sky made successively combine the advantages of both.

Faint-Object Camera. Let us only mention the faint-object camera (FOC) that is a versatile imaging instrument, essentially devoted to investigations involving very faint and very remote objects (42). Two receivers that are imaging detectors work in a photon-counting mode. They are placed at the foci of two optical relays that correct for the residual astigmatism and field curvature of the ensemble telescope, COSTAR. They convert the $f/24$ telescope beam into $f/48$ and $f/96$ beams, giving pixel sizes, respectively, of 44 and 22 mas, compared with the theoretical angular resolution of 66 mas at 633 nm. So, to exploit the full resolution capability, a facility for imaging at $f/288$ can be inserted in the $f/96$ path, giving a $7.5'' \times 7.5''$ field of view. Several additional instrumentats can also be inserted in the beams, such as various filters, an objective prism, a polarizer, a coronagraph, and a long-slit spectrograph. So, in the program of work of the FOC, astrometry is just one of the many techniques to investigate a very small field and is not a primary objective for its own sake.

Projects for the Future

The successful entry of astrometry among space astronomical techniques is a powerful incentive to devise and propose to space agencies new more powerful, more effective space astrometric missions with a better science return to cost ratio. Many projects have been presented during the last decade. Some have already been thoroughly studied, and engineered descriptions of a possible realization exist. Other are in a more dormant state. In this section, the two most ambitious projects are described. However, they have not yet been built, so the information given here is provisional. Other projects will only be mentioned, even if some might be launched before those more sophisticated.

The Space Interferometry Mission. The primary objective of the Space Interferometry Mission (SIM), scheduled for launch by NASA in 2005, is to measure stellar distances via parallaxes and apparent proper motions to discover small perturbations of their motions that could be interpreted as caused by planets, in particular Earth-sized planets. A second objective is to measure large angles and to construct a rigid grid of star positions that covers the whole sky, as the basis of a global celestial reference frame.

Description of the Instrument. The principle is that of a Michelson phase interferometer already in use on the ground and in a configuration now tested in actual size at Mount Palomar Observatory. The Palomar Testbed Interferometer is now operational and regularly observes some 100 stars per night by remote control from the JPL. Two options of the spacecraft have been studied. The first involved seven siderostats arranged linearly on a 10-meter boom (43). We describe a second one, the so-called RainBird configuration which will probably be chosen for flight. It consists of two collector pads placed symmetrically on a 7-meter, high-precision rail with respect to the combiner pad (Fig. 15). An ensemble of solar arrays and sun shades is placed on the end of a boom to protect the instrument continually from direct sunshine.

Principle of Measurements. A sketch of the instrument is given in Fig. 16. The two collectors receive light from a star and send it to a beam combiner. One of the beams is directed into a controlled delay line so that the

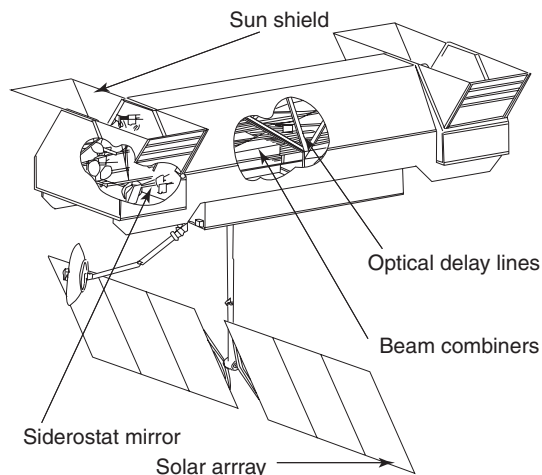


Figure 15. Oblique deployed view of the Space Interferometry Mission.

external path delay is equal to the internal one. Thus, interferometric fringes are obtained at the detector. The delay line is activated so that the central fringe remains on the central line of the detector. The reading of the delay line added to the calibrated other internal paths gives the path delay x which is recorded. If D is the baseline, also calibrated internally,

$$x = B \cos \theta, \quad (20)$$

where θ is the angle between the baseline and the direction of the star. Two more interferometers between articulated light collectors on the same baseline measure similarly, the direction of two bright stars with known positions, part of the

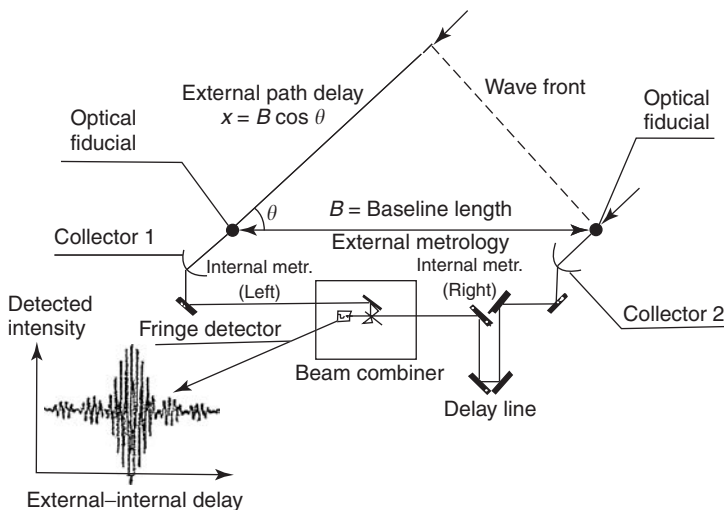


Figure 16. Principle of the SIM interferometer. The peak of the interference pattern occurs when the internal path equals the external path delay.

global rigid grid. From the results, one deduces the space orientation of the baseline at an accuracy of 30 mas. Associated with the main observation, θ defines a portion of a celestial small circle around the baseline on which the star is located. Observations at different orientations of the baseline provide the position of the star at the intersection of the loci.

Calibrations and Expected Results. The objective is to achieve microsecond of arc (μas) astrometry at the end of the 5-year mission. During this time, the spacecraft, which is not on a geocentric orbit, progressively moves away from Earth and reaches finally a distance of 95 million kilometers. This choice minimizes the speed of variation of the aberration which is quite important for an Earth satellite. To perform this correction with a superabundant accuracy of $0.7\mu\text{as}$, it will be necessary to know the velocity of the spacecraft to an accuracy of 4 mm s^{-1} . The velocity will be determined from tracking by the NASA Deep Space Network observing two hours a day.

But the major technical challenges are the onboard measurements of distances between various fiducial corner cubes needed to determine internal delays. There are two types of measurements. First, absolute determinations of distance up to 12 meters are to be made with an accuracy of $10\mu\text{m}$. This condition is not so stringent, but it requires a very rigorous stability of the lasers. On the other hand, relative metrology concerned with the variations of the baselines must be accurate to 1 or 2 picometers to achieve the astrometric objectives, that is, 10 million times better than the absolute length measurements. To achieve this, several different methods have been proposed and tested in the laboratory (44,45). The conclusion is that such measurements are possible on the ground and also in space. In all cases, the corresponding calibration cycles will be performed every hour for the relative internal measurement needs and every few days for the external delay. In the error budget, one has to take into account, in addition, thermal effects, even if they are reduced by severe thermal control, fringe measurement errors, and beam walk error produced by mispointing the compressed beams, warping of the pads, shear of the metrology beams, etc.

When these calibrations are performed, it is expected that a $7.5\mu\text{as}$ precision measurement of one locus of the star position may be obtained in 0.2 s for stars of magnitude 8, 10 s at magnitude 12, 7 minutes at magnitude 16, and four and a half hours at magnitude 20. The uncertainty also decreases as the inverse of the square root of the exposure time, so that brighter stars may be observed longer without scheduling consequences, because it is essentially the slow (0.25° per second) pointing motion and acquisition time that will limit the scheduling. Finally, an accuracy of $1\mu\text{as}$ will be achievable for a majority of the 10,000 stars expected to be on the program, and $4\mu\text{as}$ for the global grid of 4000 stars.

GAIA. The Global Astrometry Instrument for Astrophysics (GAIA) is a spacecraft proposed to the European Space Agency (ESA) as a follow-up to Hipparcos. It is not yet an approved project. If it was to be programmed by ESA before the year 2001, it could be launched in 2008–2009. As in the case of SIM, several successive versions of the project were studied. Originally, the letter I of GAIA stood for Interferometry because it was proposed that the receiver would be a Fizeau interferometer that produces fringes that would give a more precise measurement (46,47). But engineering studies, financed by ESA, proved that if one replaces the two interferometric apertures by a single mirror encompassing

them, the same accuracy of measurements results with a tremendous gain in limiting luminosity and a smaller overall cost. As it is designed now (48), GAIA is a versatile all-sky survey instrument which, in addition to performing very accurate astrometry, will do multicolor photometry, radial velocity measurements, and some narrow-band photometry for all stars up to magnitude 17. The highest priority is, however, astrometry, and the objective is to get astrometric measurements up to magnitude 20, that is, on more than a billion stars.

Description of the Payload. The principle of GAIA is identical to Hipparcos in the sense that two fields of view separated by a basic angle (here, $\gamma = 106^\circ$) are simultaneously observed. However, rather than directing the two fields of view on the same focal surface, there are two separate identical telescopes; each has its own receiving subsystem. The invariance of the basic angle is monitored by laser interferometers. The layout of the two full-reflective, three-mirror telescopes, thermally controlled at a temperature below 200 K, is shown in Fig. 17. The space left between the two telescopes is filled by a third telescope adapted for radial velocity measurements and spectrophotometry. The primary mirror of the astrometric telescopes is a rectangle of 1.4×0.5 meters. The optics give an equivalent 50-meter focal length so that the useful field of view of $0.66^\circ \times 0.66^\circ$ is projected on a detector whose dimensions are 575×700 mm. The satellite rotates around an axis perpendicular to the telescope layout and scans the sky following a law analogous to that of Hipparcos. As in the case of SIM, it was recognized that the observations should be made far from Earth. For GAIA, the choice is a Lissajous type of orbit around the Laplace L2 point of the Sun-Earth system.

The Detector System. The detector (Fig. 18), placed on the focal surface of the telescope, includes 250 CCD arrays of 2100×2660 pixels organized in 10 along-scan (horizontal) strips. Each array is 24 mm wide along scan and 57 mm long in the vertical direction. The pixel size along scan is $9 \mu\text{m}$ ($36 \mu\text{as}$ in the sky) $\times 27 \mu\text{m}$ ($108 \mu\text{as}$), compatible with the shape of the point-spread function. The observing strategy is the scan mode already mentioned earlier. The transfer

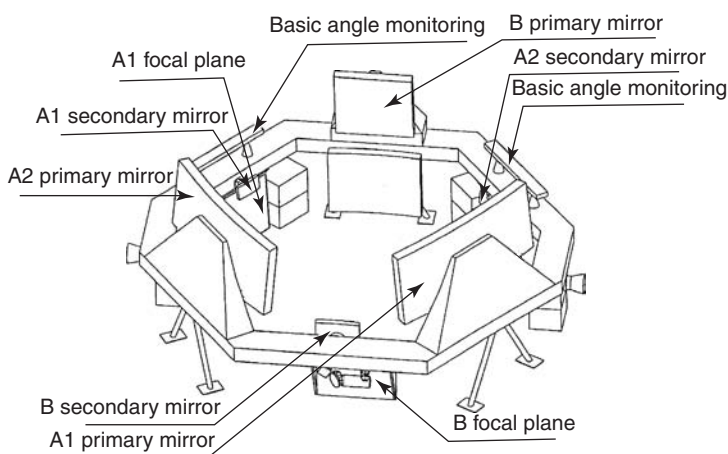


Figure 17. Layout of the three telescopes of the GAIA project. A1 and A2 are the astrometric telescopes; B is the radial velocity/photometry telescope.

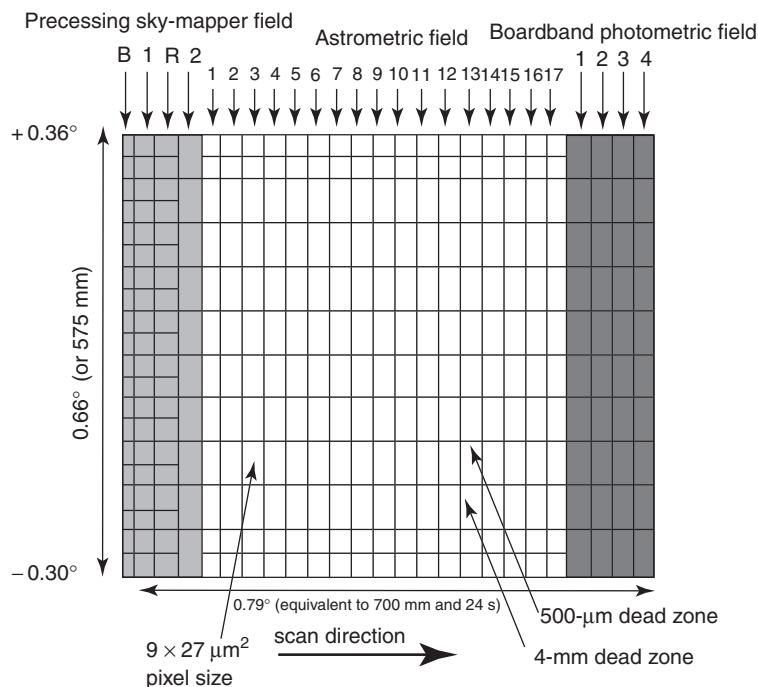


Figure 18. Arrangement of the CCDs on the GAIA focal surface.

of charges is done continuously, and the speed corresponds to the rate of rotation of the instrument. The collection of charges is done in the 4-mm dead zones between the arrays.

The vertical columns of CCDs do not have all the same functions. The first four in the scan direction form the star-mapper. All signals are processed. Those corresponding to an input catalog are used to control the attitude, but all are analyzed, and the position of those that are sufficiently bright to be measured are used to sort the useful signals in the astrometric CCDs. The latter occupy the next 17 columns in which the astrometric data are acquired. The last four columns are used for broadband photometry in different colors.

Because of the curvature of the focal surface, the arrays will be slightly tilted and individually sequenced to compensate for optical distortion. Each individual CCD features a special operating mode, which may or may not be activated, that allows reducing the integration time and acquiring bright stars with no saturation which occurs for magnitudes brighter than 12.

Expected Results. The astrometric performance depends on magnitude and also on color. Many more photoelectrons will be received from red stars than from blue ones. For intermediate stars (solar type), the accuracy floor up to magnitude 12 is $3 \mu\text{as}$. It degrades for fainter stars and is $10 \mu\text{as}$ for magnitude 15. In other terms, for some 35 million stars, the astrometric accuracy in position, parallaxes, and yearly proper motions will be better than $10 \mu\text{as}$. For magnitude 16, the numbers will be, respectively, 60 million stars and $18 \mu\text{as}$; for magnitude 18; 300 million stars and $55 \mu\text{as}$; and for magnitude 20, more than a billion stars

and 0.2 mas. So even at the limiting magnitude of 20, it is expected that GAIA will be five times more precise than Hipparcos for its bright stars.

For the first time, radial velocities will be systematically measured all over the sky. The wavelength interval provisionally set is 847–879 nm. The expected precision is a few km s^{-1} for stars up to magnitude 17 that have spectral lines in this interval. This means that more than one hundred million radial velocities would complete the proper motion and parallax measurements for space velocities.

Other Projects. Among various proposals presented for space astrometry, three have some chance of being realized and launched.

FAME. The Fizeau Astrometric Mapping Explorer (FAME) was conceived to meet the NASA philosophy of “small, fast, and cheap.” It consists of two dilute aperture telescopes operating in the Fizeau mode (49), a concept that was also originally proposed for GAIA (46), but on a smaller scale. The apertures would be approximately 10×20 cm and the baseline 50 cm. The angle between the telescopes would be 80° . The operations would be quite similar to Hipparcos with a predefined scanning law. The detectors will be CCD arrays working in scan mode as in GAIA. The expected limiting magnitude is 15, and the accuracies for astrometric parameters would range from $40 \mu\text{as}$ at magnitude 9 to $180 \mu\text{as}$ at magnitude 12 and 0.8 mas for the faintest stars.

DIVA. The Deutsches Interferometer für Vielkanalphotometrie und Astrometrie (DIVA) was submitted to the German Space Agency and is designed to perform astrometric and photometric observations of a million stars (50). The configuration consists of two Fizeau interferometers with a baseline of 10 cm; the receiver is a mosaic of CCDs. The operations would be analogous to Hipparcos. It is expected that the resulting catalog will be complete to magnitude 10.5 with an accuracy of astrometric parameters of the order of 0.3 mas.

LIGHT. The Light Interferometer for the study of Galactic Halo Tracers (LIGHT) is much more ambitious than the two preceding projects. It is a Japanese project for global astrometry (51). It is composed of four sets of Fizeau interferometers with a beam combiner unit of 1 meter baseline. The operations are again similar to Hipparcos. The prime scientific objective, as witnessed by its name, is to monitor parallaxes and proper motions of distant galactic and halo giant red stars up to visual magnitude 18, but all stars in this magnitude range will also be observed. The expected accuracy of the astrometric parameters is $50 \mu\text{as}$ and should involve some hundred million stars. The satellite would also include photometry in visual and near-infrared bands.

BIBLIOGRAPHY

1. Arias, E.F., P. Charlot, et al. *Astron. Astrophys.* 303: 604 (1995).
2. Ma, C., E.F. Arias, et al. *Astron. J.* 117: 516 (1998).
3. Soffel, M.H. *Relativity in Astrometry, Celestial Mechanics and Geodesy*. Springer-Verlag, Berlin, 1989, pp. 63–65.
4. Van de Kamp, P. *Principles of Astrometry*. W.H. Freeman, San Francisco, 1967, pp. 165–175.
5. Green, R.M. *Spherical Astronomy*. Cambridge University Press, Cambridge, 1985, pp. 478–485.

6. Perryman, M.A.C., A. Brown, G., A., et al. *Astron. Astrophys.* 331: 81 (1998).
7. Kovalevsky, J. *Modern Astrometry*. Springer-Verlag, Berlin, 1995.
8. Monet, D.G., C.C. Dahn, et al. *Astron. J.* 103: 638 (1992).
9. Baldwin, J.E., M.J. Beckett, et al. *Astron. Astrophys.* 306: L13 (1996).
10. Hartkopf, W.I., B.D. Mason, and H.A. McAlister. *Astron. J.* 108: 370 (1996).
11. Hummel, C.A., D. Mozurkewich, et al. *Astron. J.* 108: 326 (1996).
12. Fricke, W., Schwan, et al. *Fifth Fundamental Catalogue (FK5)*. Veroff. Astron. Rechen-Institut Heidelberg 32, (1988).
13. Lacroute, P., *Trans. IAU XIII*: 63 (1967).
14. ESA, *The Hipparcos and Tycho Catalogues*. ESA Publication Division, ESTEC, Noordwijk, SP-1200, 1997.
15. Turon, C., and 53 other authors, *The Hipparcos Input Catalogue*. ESA Publication Division, ESTEC, Noordwijk, SP-1136, 1992.
16. Kovalevsky, J., J.-L. Falin, et al. *Astron. Astrophys.* 258: 7 (1992).
17. Lindegren, L., E. Høg, et al. *Astron. Astrophys.* 258: 18 (1992).
18. Van Leeuwen, F. *Space Sci. Rev.* 81: 201 (1997).
19. Donati, F., M. Froeschlé, et al. *Astron. Astrophys.* 258: 41 (1992).
20. van Leeuwen, F., M.J. Penston, et al. *Astron. Astrophys.* 258: 53 (1992).
21. van der Mrel, H., and C. Petersen. *Astron. Astrophys.* 258: 60 (1992).
22. Bernstein, H.-H. *Hipparcos Venice'97*, M.A.C. Perryman and P.L. Bernacca (eds), ESA Publication Division, ESTEC, Noordwijk, SP-402, 1997, pp. 705–708.
23. Kovalevsky, J., L. Lindegren, et al. *Astron. Astrophys.* 323: 620 (1997).
24. Arenou, F., L. Lindegren, et al. *Astron. Astrophys.* 304: 52 (1995).
25. Mignard, F., S. Söderhjelm, et al. *Astron. Astrophys.* 304: 82 (1995).
26. Gazengel, F., A. Spagna, et al. *Astron. Astrophys.* 304: 105 (1995).
27. Perryman, M.A.C. and P.-L. Bernacca (eds). *Hipparcos Venice'97*, ESA Publication Division, ESTEC, Noordwijk, SP-402, 1997.
28. Kovalevsky, J. *Annu. Rev. Astron. Astrophys.* 36: 99 (1998).
29. Mignard, F., M. Froeschlé, and J.-L. Falin. *Astron. Astrophys.* 258: 142 (1992).
30. Evans, D.W., F. van Leeuwen, et al. *Astron. Astrophys.* 258: 149 (1992).
31. Høg, E., U. Bastian, et al. *Astron. Astrophys.* 258: 177 (1992).
32. Halbwachs, J.-L., E. Høg, et al. *Astron. Astrophys.* 258: 193 (1992).
33. Høg, E., U. Bastian, et al. *Astron. Astrophys.* 258: 201 (1992).
34. Grossmann, V., G. Bässgen, et al. *Astron. Astrophys.* 304: 110 (1995).
35. Hall, D.N.B. *The Space Telescope Observatory*. NASA, Washington, DC, 1982.
36. Burrows, C.J., J.A. Holtzman, et al. *Astrophys. J. Lett.* 369: L21 (1991).
37. Lasker, B.M., H. Jenker, J.L. Russell, *Mapping the Sky*. S. Debarbat, et al. (eds), IAU Symposium 133, Kluwer Academic, Dordrecht, 1988, pp. 229–233.
38. O'Dell, C.R., *The Space Telescope Observatory*, D.N.B. Hall (ed.), NASA, Washington, DC, 1982, p. 20.
39. Duncombe, R.L., W.H. Jefferys, et al. *Adv. Space Res.* 11 (2): 87 (1991).
40. Taff, L.G. *Adv. Space Res.* 11 (2): 97 (1991).
41. Seidelmann, P.K. *Adv. Space Res.* 11 (2): 103 (1991).
42. Macchetto, F. *The Space Telescope Observatory*, D.N.B. Hall (ed.), NASA, Washington, DC, 40, 1982, p. 40.
43. Boden, A., S. Unwin, and M. Shao. *Hipparcos Venice'97*, M.A.C. Perryman, and P.L. Bernacca (eds), ESA Publication Division, ESTEC, Noordwijk, SP-402, 1997, pp. 789–792.
44. Gursel, Y. *Proc. SPIE* 1947: 188 (1993).
45. Gursel, Y. *Proc. SPIE* 2477: 240 (1995).
46. Lindegren, L., and M.A.C. Perryman. *Future Possibilities for Astrometry in Space*. ESA Publication Division, ESTEC, Noordwijk, SP-379, 1997, pp. 23–32.

47. Høg, E., U. Bastian, and W. Seifert. *Hipparcos Venice'97*, M.A.C. Perryman and P.L. Bernacca (eds), ESA Publication Division, ESTEC, Noordwijk, SP-402, 1997, pp. 783–788.
48. Perryman, M.A.C. *Int. Fed. Astronaut. Symp.*, Melbourne, 1998.
49. Seidelmann, P.K., K.J. Johnston, et al. *Future Possibilities for Astrometry in Space*. ESA Publication Division, ESTEC, Noordwijk, SP-379, 1997, pp. 187–189.
50. Röser, S., U. Bastian, et al. *Hipparcos Venice'97*, M.A.C. Perryman and P.L. Bernacca (eds), ESA Publication Division, ESTEC, Noordwijk, SP-402, 1997, pp. 777–782.
51. Yoshisawa, M., K. Sato, et al. *Hipparcos Venice'97*, M.A.C. Perryman, and P.L. Bernacca (eds), ESA Publication Division, ESTEC, Noordwijk, SP-402, 1997, pp. 795–797.

READING LIST

- Eichhorn, H. *Astronomy of Star Positions*. Frederic Ungar, New York, 1974.
Green, R.M. *Spherical Astronomy*. Cambridge University Press, Cambridge, 1985.
Hall, D.N.B. *The Space Telescope Observatory*, NASA, Washington, DC, 1982.
Kovalevsky, J. *Modern Astrometry*. Springer-Verlag, Berlin, 1995.
Soffel, M.H. *Relativity in Astrometry, Celestial Mechanics and Geodesy*. Springer-Verlag, Berlin, 1989.
Van de Kamp, P. *Principles of Astrometry*. W.H. Freeman, San Francisco, 1967.
Van de Kamp, P. *Stellar Paths*. Reidel, Dordrecht, 1981.

JEAN KOVALEVSKY

Cerga-Observatoire de la Côte d'Azur
Grasse, France

P

PATHFINDER MISSION TO MARS

On July 4, 1997, Mars *Pathfinder* landed safely on the surface of Mars. Designed under the new “faster, cheaper, and better” *Discovery* program philosophy, the lander deployed and navigated a small rover named “*Sojourner*” onto the Ares Valles landing site and began collecting data from its onboard scientific instruments. Designed primarily as an entry, descent and landing demonstration, *Pathfinder* returned 2.3 billion bits of new data, including over 17,000 images, 16 chemical analyses of rocks and soil, and 8.5 million individual temperature, pressure and wind measurements. *Sojourner* traversed approximately 100 meters (330 feet) clockwise around the lander exploring about 200 square meters (2,153 square feet) of area. See Fig. 1. The mission captured the imagination of the public, garnered front page headlines during the first week of mission operations, and went on to become one of NASA’s most popular missions. A total of about 56.6 million people visited the *Pathfinder* Web Pages during the first month of the mission, with 4.7 million people visiting the Web Pages on July 8, 1997 alone, making the *Pathfinder* landing by far the largest Internet event in history up to that time.

Mission Summary. The Mars *Pathfinder* mission was the second mission launched under the National Aeronautics and Space Administration’s (NASA) *Discovery* Program. The *Discovery* missions were developed for small planetary missions with a maximum three-year development cycle and a cost cap of \$150 million (Fiscal Year 1992) for development that focused on engineering, science, and technology objectives. Originally conceived as an engineering demonstration of key technologies and concepts for use in future missions to Mars, the primary objective was to demonstrate a low-cost cruise, entry, descent, and landing system that could safely place a variety of science instruments on the surface of Mars. For *Pathfinder*, the cost of the mission was \$171 million (Fiscal Year 1996),

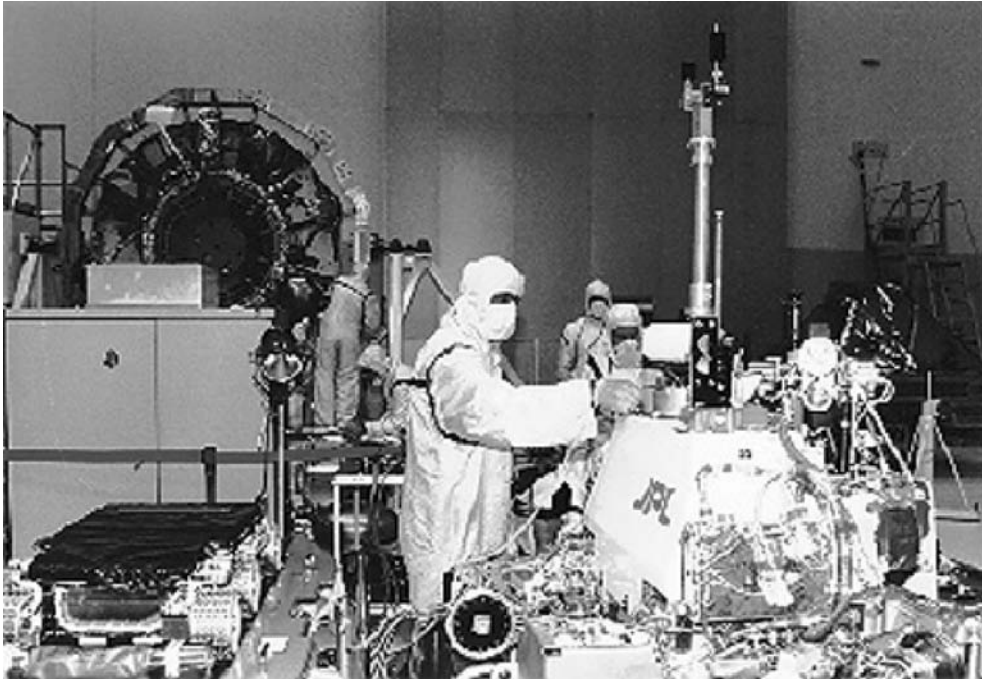


Figure 1. Mars *Pathfinder*, rover, and cruise stage being unpacked at the Kennedy Space Center. (Image courtesy of KSC/NASA.)

the *Delta II* launch vehicle was an additional \$55 million, the development and operations of the rover cost an additional \$25 million, and \$14 million was allotted for operations.

Mission and Spacecraft Overview

The *Pathfinder* spacecraft or flight system consisted of three major components: the cruise stage, the entry decent subsystem, and the lander, which consisted of the science instruments and the rover.

Cruise Phase. The cruise phase of the Mars *Pathfinder* mission began with the successful launch atop a *Delta II* rocket from the Kennedy Space Center in Florida on December 4, 1996. See Fig. 2. Once in earth orbit, the spacecraft was given a final boost with the help of a solid-fuel rocket motor called a *Payload Assist Module (PAM-D)*. This ‘kick-stage’ gave the spacecraft just the right amount of velocity increase it needed to escape Earth’s gravity and enter its own orbit around the Sun. Once spent, the third stage was jettisoned.

At separation from the upper stage, the spacecraft was in Earth’s shadow and spinning at 20rpm. An onboard sequence of events was activated once the separation microswitch detected the separation. The Deep Space Network (DSN) initiated spacecraft acquisition and lockup activities using a 34-meter (112-foot) antenna located in the California desert. (See Deep Space Network,



Figure 2. Mars *Pathfinder* launch onboard a Delta II on December 4, 1996. (Image Courtesy of KSC/NASA.)

Evolution of Technology). As soon as acquisition occurred, the engineering telemetry broadcast by the spacecraft was received on the ground at a rate of 40 bits per second (b/s). This telemetry consisted of a combination of real time engineering data and stored data from launch, separation, and Earth/Sun acquisition. See Fig. 3.

The spacecraft automatically determines its orientation in space by first determining the location of the Sun with respect to the spin axis of the spacecraft using a Sun sensor located on the top of the cruise stage. This procedure, known as Sun acquisition, was supposed to provide the spacecraft with the information it needed to reduce the spin rate from 20 rpm to a nominal 2 rpm. But due to some difficulties during launch, it was soon discovered that two of the five sensors had been damaged with an unknown, foreign substance. A software patch was developed which corrected the problem and by using the data from the three working sensors, engineers were able to slow the spacecraft down. Once the spacecraft had cleared the Moon's orbit and safely spun down to 2 rpm, the star scanner was activated. After star identification had been confirmed, the *Attitude and Information Management (AIM)* computer calculated the spacecraft's orientation and position, and started its seven-month trip to Mars.

Mars *Pathfinder* used an Earth-Mars transfer orbit. The total flight time from Earth to Mars took seven months. See Fig. 4 for a view of the interplanetary trajectory, as it would look from above the Sun. During the seven-month cruise to Mars, a number of activities were performed to maintain the health of the entry vehicle, lander and rover. Navigation was required to maintain the flight path,

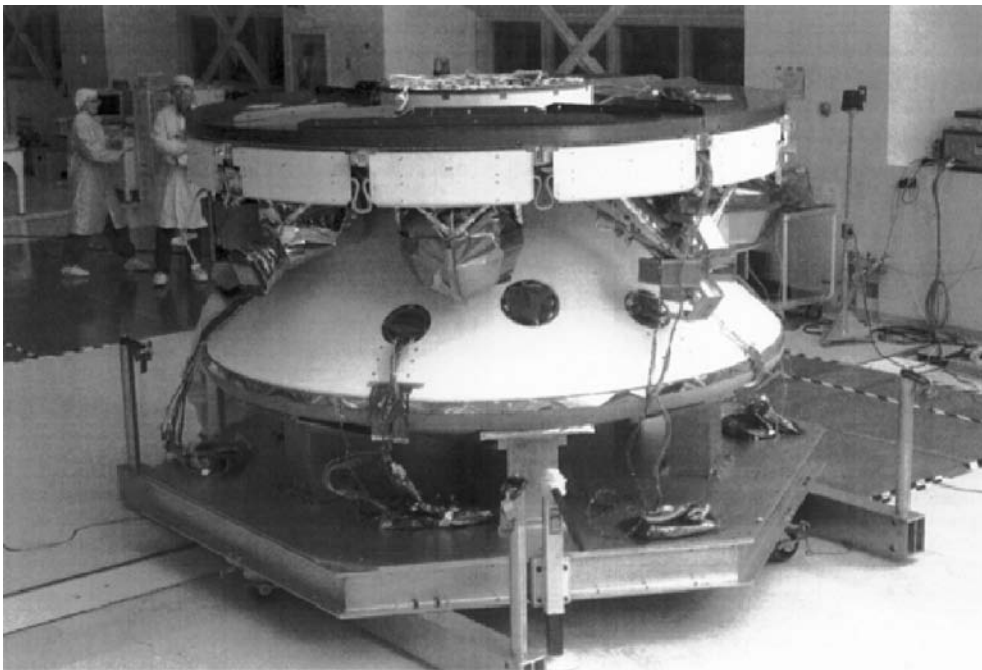


Figure 3. Mars *Pathfinder* in cruise configuration. The red panels are the solar cells that will supply power during the seven month cruise. (Image courtesy of JPL/NASA.)

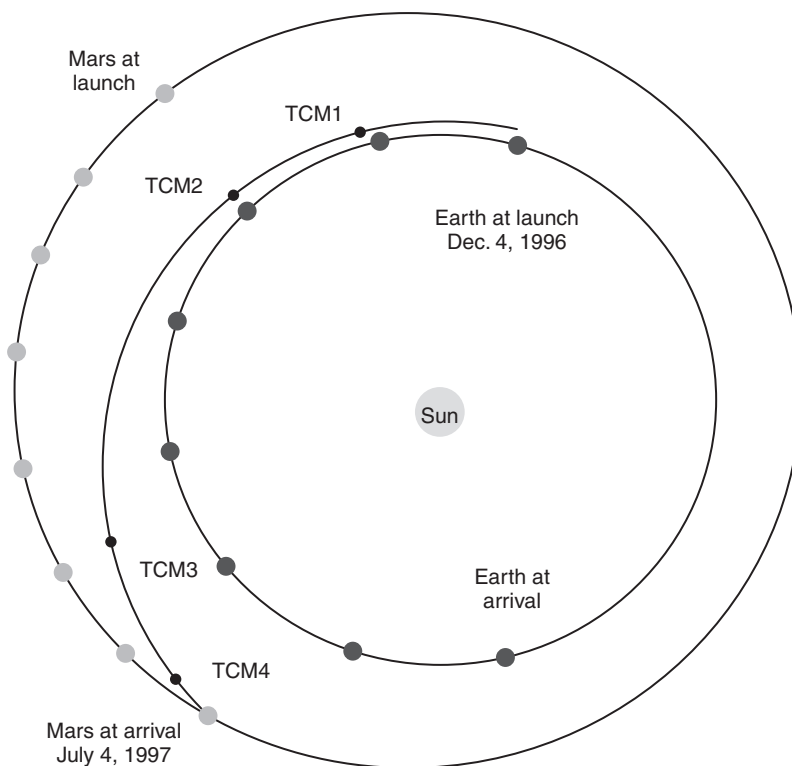


Figure 4. Mars *Pathfinder* cruise trajectory.

and the various spacecraft subsystems were monitored and adjusted as needed to keep them operating at peak efficiency.

Cruise activities began once the spacecraft was safely out of Earth orbit. After the attitude was established and the spacecraft was determined to be healthy, the flight team began a two-week initial characterization and calibration period. Systems included the solar array and battery, thermal control, attitude determination and control, and the communication subsystems. The primary spacecraft activities during the first month of cruise were to collect and downlink relevant engineering telemetry and tracking data, initial spacecraft health checks and calibrations, and attitude maneuvers to maintain the correct Earth/Sun geometry. One health check of the Rover and Science Instruments occurred on December 19, 1996.

Measurements of the spacecraft range to Earth and the rate of change of this distance were collected during every *DSN* station contact and sent to the navigation specialists of the flight team for analysis. They used this data to determine the true path the spacecraft was flying, and to determine corrective maneuvers needed to maintain the desired trajectory. The first of four *Trajectory Correction Maneuvers (TCMs)* was scheduled on January 4, 1997 to correct any errors collected from launch. The magnitude of this maneuver was less than 75 meters (246 feet) per second. Navigation was an ongoing activity that continued until the spacecraft entered the atmosphere of Mars on the 4th of July.

After TCM-1, the flight team transitioned from a “spacecraft checkout mode” to a more routine “spacecraft monitoring mode”. DSN tracking coverage was reduced from three contacts a day to three per week to allow other spacecraft like Mars *Global Surveyor* and *Galileo* to use the DSN time. Spacecraft health and performance telemetry was downlinked at 40 b/s or greater during each tracking pass.

A key activity that took place during cruise was the designing and building of command sequences that dictated to the spacecraft how it was to perform each of the activities required. Each cruise command sequence was generated and tested, and then uplinked approximately once every four weeks during one of the regularly scheduled DSN passes. The uplink generation process required 14 days for planning, sequence generation, verification, and commanding.

Two more trajectory correction maneuvers were performed in early February and early May to further reduce any navigation guidance errors. TCM-2 required less than 10 meters (33 feet) per second, and TCM-3 was smaller still, less than 1 meter (3.3 feet) per second. These two maneuvers further reduced any guidance error detected from navigation measurements during cruise.

Starting 45 days prior to entry, tracking was increased to three passes a day and the flight team stepped up its preparation for atmosphere entry and landing. A final health and status check of the instruments and rover was performed on June 4, 1997. A fourth and final trim maneuver was performed on June 24, requiring less than 0.50 meters (1.65 feet) per second due to the accuracy of the previous maneuvers. On June 30, the spacecraft performed a turn to the entry attitude, where it remained until atmosphere entry. The roll thrusters increased the spacecraft spin rate from 2 to 10 rpm for entry. At that time, the cruise phase ended and the flight team transitioned to the entry, landing, and surface operations phases. See Fig. 5.

During the final day of approach, the navigation team produced orbit solutions on a regular basis, and adjustments were made to the computer programs that determine when the parachute should be deployed. At 6 hours out, the final adjustments were made, and the flight team made final preparations for atmosphere entry.

Entry, Decent, and Landing Phase. The fast-paced approach of *Pathfinder* to Mars began with venting of the heat rejection system’s cooling fluid about 90 minutes prior to landing. See Fig. 6. This fluid is circulated around the cruise stage perimeter and into the lander to keep the lander and rover cool during the seven month cruise phase of the mission. Its mission fulfilled, the cruise stage was then jettisoned from the entry vehicle about one-half hour prior to landing at a distance of 8,500 kilometers (5,100 miles) from the surface of Mars. Several minutes before landing, the spacecraft began to enter the outer fringes of the atmosphere about 125 kilometers (75 miles) above the surface. Spin stabilized at 2 rpm, and traveling at 7.5 kilometers (4.5 miles) per second the vehicle entered the atmosphere at a shallow 14.8-degree angle. A shallower entry angle would result in the vehicle skipping off the atmosphere, while a steeper entry would not provide sufficient time to accomplish all of the entry, descent and landing tasks. A *Viking* -derived aeroshell (including the heatshield) protected the lander from the intense heat of entry. At the point of peak heating the heatshield absorbed more than 100 megawatts of thermal energy. The Martian atmosphere slowed

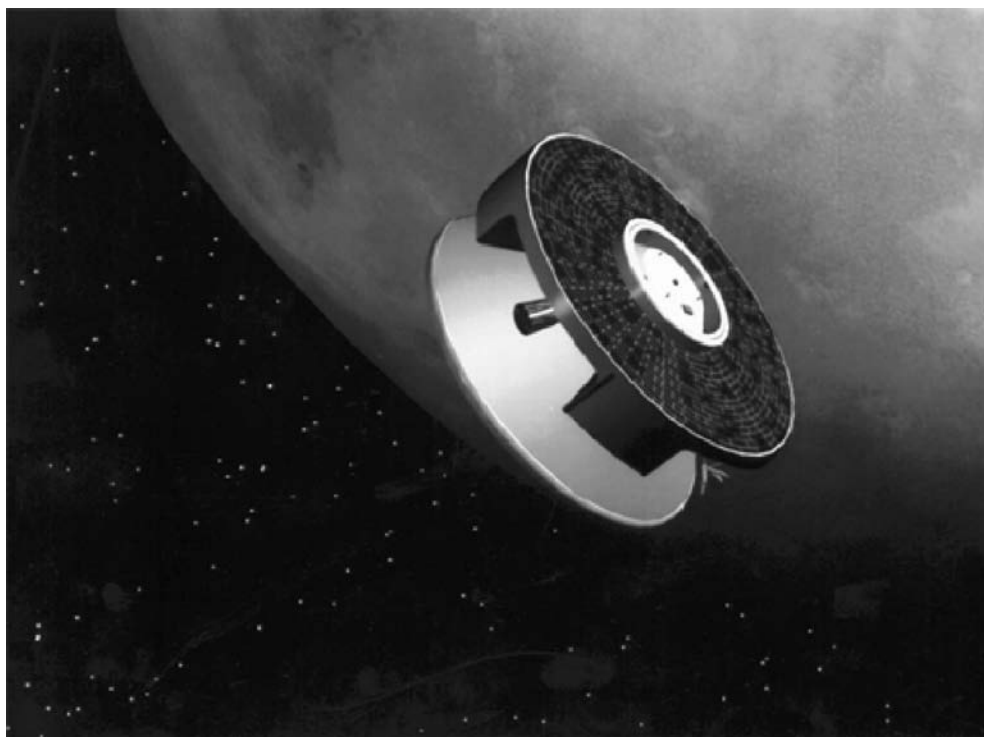


Figure 5. Artist renditions of Mars *Pathfinder* as it enters the Martian atmosphere. (Image courtesy of JPL/NASA.)

the vehicle from 7.5 kilometers per second to only 400 meters per second (900 miles per hour).

Entry deceleration of up to 20 gees, detected by on-board accelerometers, set in motion a sequence of preprogrammed events that are completed in relatively quick succession. Deployment of the single, 8-meter (24-foot) diameter parachute occurred 2 minutes and 14 seconds after atmospheric entry at an altitude of 9.4 kilometers (6 miles) above the surface. The parachute was similar in design to those used for the *Viking* program but had a wider band around the perimeter, which helped to minimize swinging.

The heatshield was pyrotechnically separated from the lander 20 seconds later and dropped away. See Fig. 7. The lander soon begins to separate from the backshell and “rappels” down a metal tape on a centrifugal braking system built into one of the lander petals.

The slow descent down the metal tape places the lander into position at the end of a braided Kevlar tether, or bridle, without off-loading the parachute or placing excessive loads on the backshell. The 20-meter (66-foot) bridle provides space for airbag deployment, distance from the solid rocket motor exhaust stream and increased stability. Once the lander was lowered into position at the end of the bridle, the radar altimeter was activated and began a timing sequence for airbag inflation, backshell rocket firing and the cutting of the Kevlar bridle.

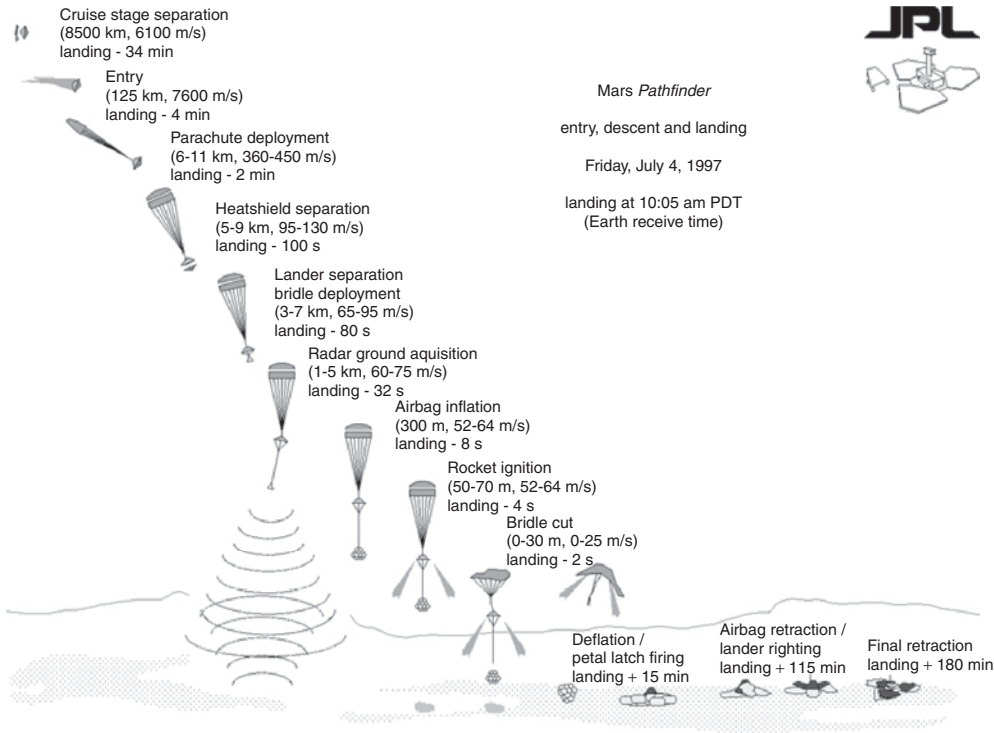


Figure 6. Entry, decent, and landing schematic for July 4, 1997. (Image courtesy of JPL/NASA.)

The lander's Honeywell radar altimeter acquired the surface about 28 seconds prior to landing at an altitude of about 1.6 kilometers (1 mile). The airbags were inflated 18 seconds later before landing at an altitude of 355 meters (less than 1/4 of a mile) above the surface. See Fig. 8. The airbags had two pyro firings, the first of which cut the tie cords and loosened the bags. The second firing, 0.25 seconds later, and 4 seconds before the rockets fired, ignited three gas generators that inflated the three 5.2-meter (17-foot) diameter bags to a little less than 1 psi. in less than 0.3 seconds.

The conical backshell above the lander contains three solid rocket motors each providing about a ton of force for over 2 seconds. The computer in the lander activates them. Electrical wires that run up the bridle close relays in the backshell which ignite the three rockets at the same instant.

The brief firing of the solid rocket motors at an altitude of 98 meters (323 feet) was intended to essentially bring the downward movement of the lander to a halt some 12 meters, ± 10 meters (40 feet, ± 30 feet) above the surface. In reality, the rockets fired approximately 21 meters (69 feet) above the surface. The bridle separating the lander and heatshield were then cut from the lander, resulting in the backshell driving up and into the parachute under the residual impulse of the rockets, while the lander, encased in airbags, fell to the surface. See Fig. 9.

Because it was possible that the backshell could be at a small angle at the moment that the rockets fire, the rocket impulse imparted a large lateral velocity

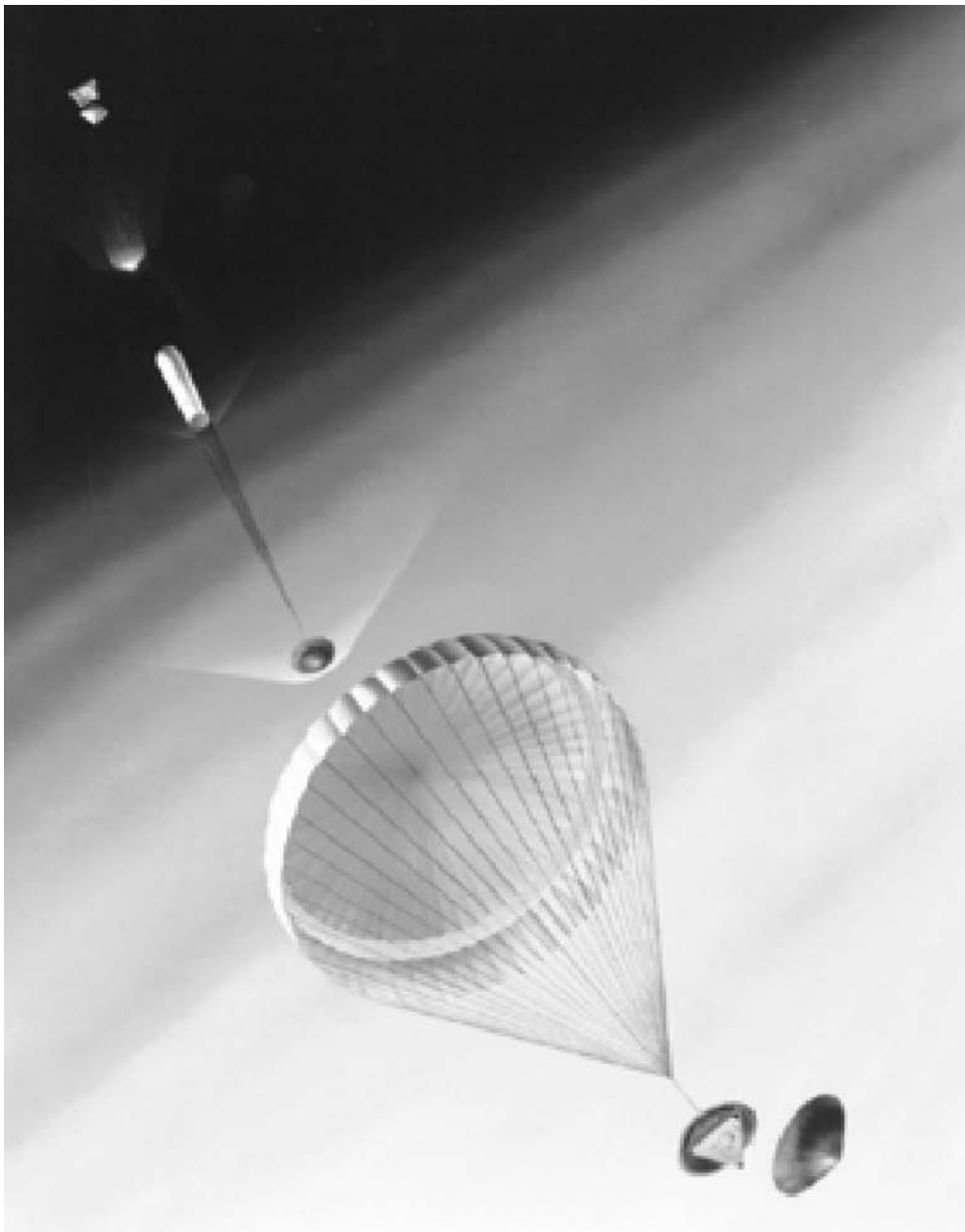


Figure 7. Artist rendition of *Pathfinder* entering the Martian atmosphere. (Image courtesy of JPL/NASA.)

to the lander/airbag combination. In fact the impact could have been as high as 25 meters per second (56 mph) at a 30-degree grazing angle with the terrain. It was expected that the lander could have bounced at least 12 meters (40 feet) above the ground and soared 100 to 200 meters (330 to 660 feet) between bounces. Tests of the airbag system verified that it was capable of much higher



Figure 8. One of the many airbag test performed prior to lift-off. Each lobe of the airbags consists of six spheres, with four lobes, one for each of the pedals. The airbags total 16 feet from the ground to the top. (*Image courtesy of JPL/NASA.*)

impacts and longer bounces. In reality, an onboard instrument calculated at least 15 bounces with the first bounce up to 12 meters (40 feet) high without any airbag rupture.

Once the lander had settled on the surface, pyrotechnic devices in the lander petal latches were blown to allow the petals to be opened. The latches locking the sturdy side petals in place were necessary because of the pulling forces exerted on the lander petals by the deployed airbag system. In parallel with the petal latch release, a retraction system began slowly dragging the airbags toward the lander, breaching vent ports on the side of each bag, in the process deflating the bags through a cloth filter. The airbags were drawn toward the petals by internal lines extending between attachments within the airbags and small winches on each of the lander sides. It took about 64 minutes to deflate and fully retract the bags. See Fig. 10.

There is one high-torque motor on each of the three petal hinges. If the lander had come to rest on its side, it would have to be righted to the base petal by opening a side petal with a motor drive to place the lander in an upright position. Once upright, the other two remaining petals would have been opened.

About three hours was allotted to retract the airbags and deploy the lander petals, but on Mars the whole operation only took 87 minutes because *Pathfinder* came to rest on the base petal. At this time, the lander's X-band radio transmitter was turned off for the first time since before it was launched on December 4, 1996.

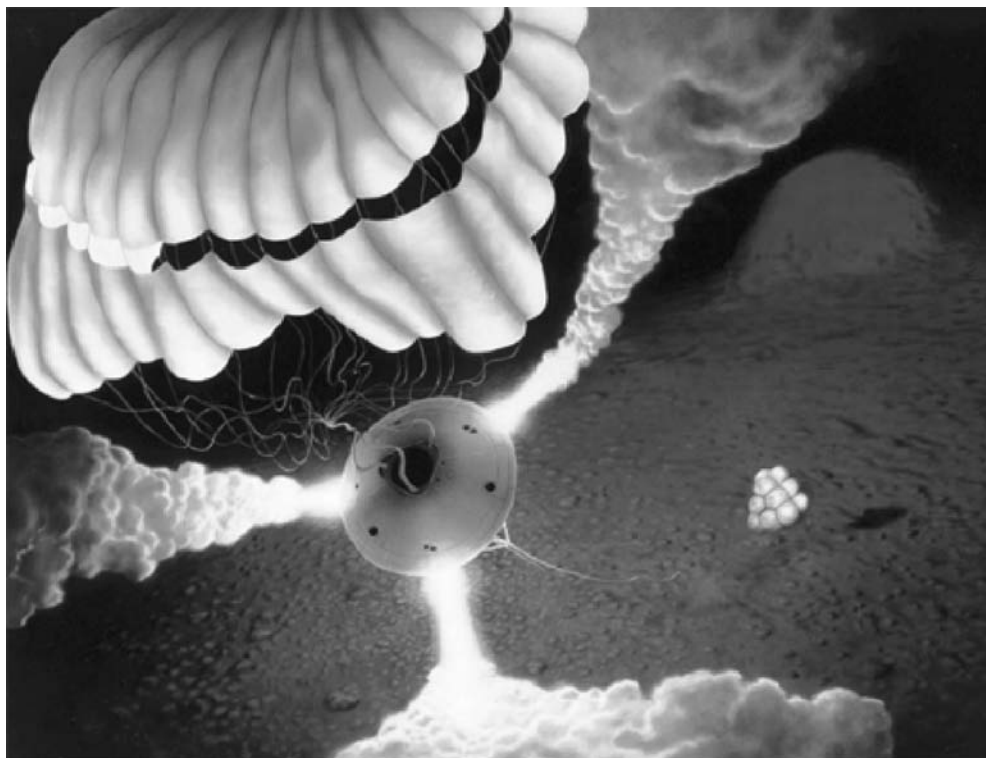


Figure 9. Artist renditions of the landing of Mars *Pathfinder* on the surface of Mars. (Image courtesy of JPL/NASA.)

This saved battery power and allowed the transmitter electronics to cool down after being warmed up during entry without the cooling system. It also allowed time for the Earth to rise well above the local horizon so that it would be in a better position for communications with the lander's low-gain antenna.

Science Instruments and Objectives

The Mars *Pathfinder* project landed a single vehicle on the surface of Mars, which included a microrover, (*Sojourner*), and several science instruments. See Fig. 11. *Sojourner's* mobility provided the capability of “ground truthing” the local landing area and investigating the surface with three additional science instruments: A stereoscopic imager with spectral filters on an extendible mast (IMP), an Alpha Proton X-Ray Spectrometer (APXS), and an Atmospheric Structure Instrument/Meteorology package (ASI/MET). See Fig. 12. These instruments allowed for investigations of the geology and surface morphology at submeter to a hundred meters scale, the geochemistry and petrology of soils and rocks, the magnetic and mechanical properties of the soil as well as the magnetic properties of the dust, a variety of atmospheric investigations and rotational and orbital dynamics of Mars.



Figure 10. Rover view of the lander on the surface of Mars. Notice how far the airbags retracted. (Image courtesy of JPL/NASA.)

Landing downstream from the mouth of a giant catastrophic outflow channel (Ares Vales) offered the potential for identifying and analyzing a wide variety of materials in the crust, from the ancient heavily cratered terrain to intermediate-aged ridged plains to reworked channel deposits. Examination of the

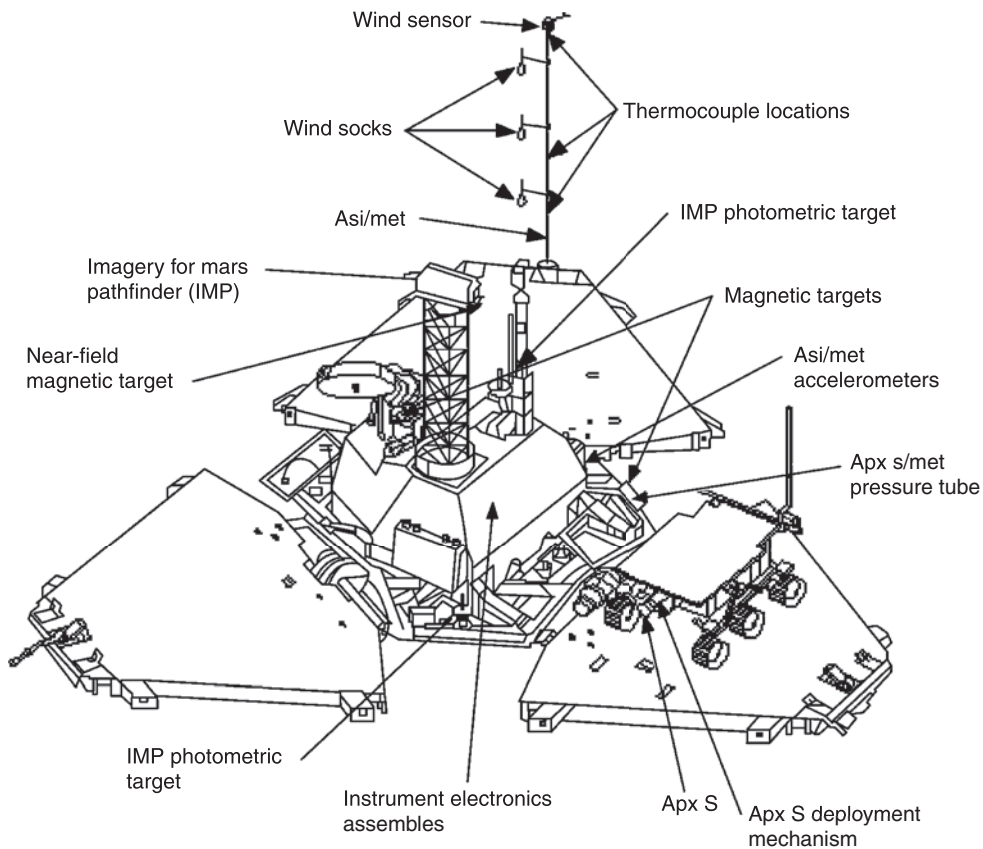


Figure 11. Computer drawing of the lander components. (Image courtesy of JPL/NASA.)

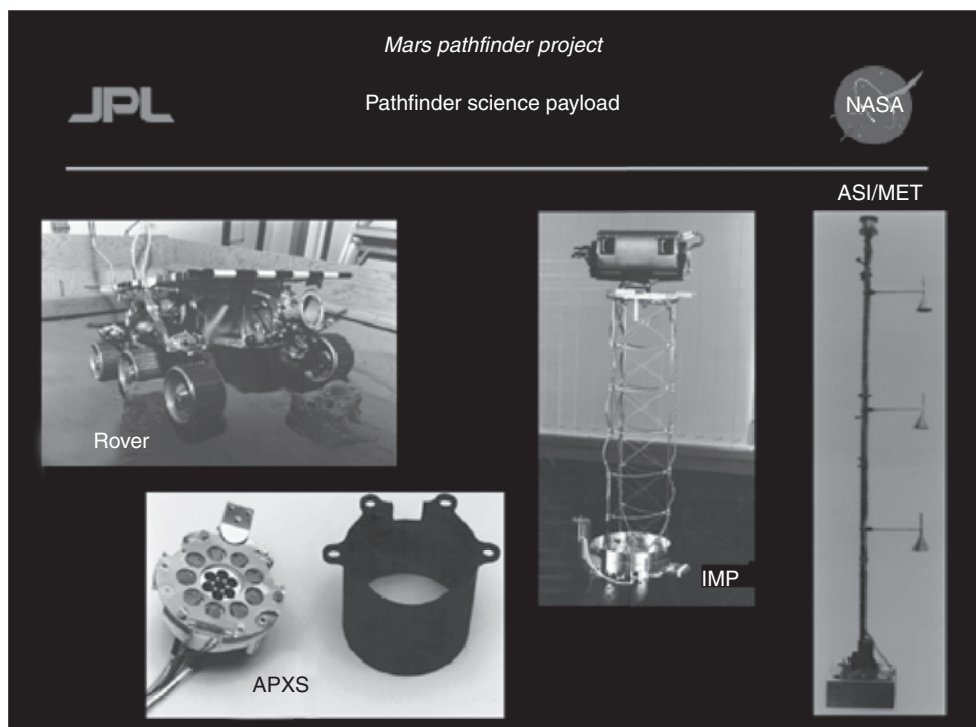


Figure 12. Mars *Pathfinder* instrument package. Imager for Mars *Pathfinder* (IMP), Alpha Proton X-Ray Spectrometer (APXS), an Atmospheric/Meteorology. (ASIMET.)

different surface materials allowed first-order scientific investigations of the early differentiation and evolution of the crust, the development of weathering products and the early environments and conditions that have existed on Mars.

Surface Morphology and Geology at Meter Scale. The Imager for Mars *Pathfinder* (IMP) examined Martian geologic processes and surface-atmosphere interactions similar to what was observed at the *Viking* landing sites. See Fig. 13. Observations of the general landscape, surface slopes and the distribution of rocks were obtained by panoramic stereo images at various times of the day. IMP was also designed to monitor any dust or sand deposition, erosion or other surface-atmosphere interactions. A basic understanding of the surface and near-surface soil properties was obtained by the rover and lander imaging of rover wheel tracks, holes dug by rover wheels, and examining any surface disruptions caused by airbag bounces or retractions.

Petrology and Geochemistry of Surface Materials. The Alpha-Proton X-Ray Spectrometer (APXS) and the visible to near-infrared spectral filters on the IMP determined the dominant elements that made up the rocks and other surface materials of the landing site. A better understanding of these materials provided answers concerning the composition of the Martian crust, and secondary weathering products (such as different types of soils). These investigations provided a calibration point for orbital remote sensing observations such as Mars

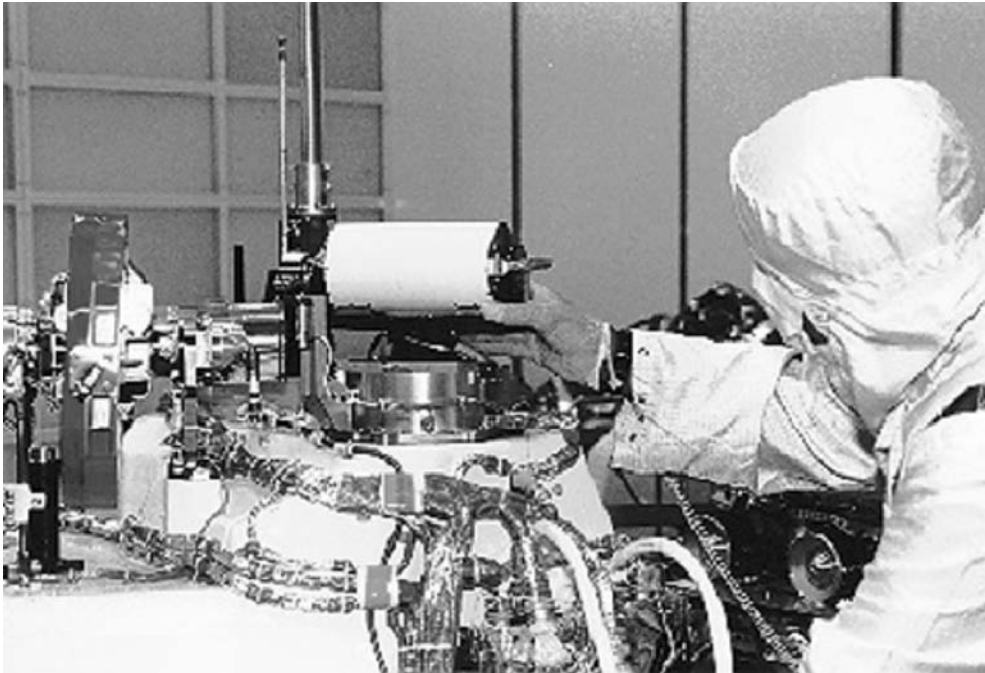


Figure 13. Figure 7 Imager for Mars *Pathfinder* (IMP) being tested before launch. (Image courtesy of KSC/NASA.)

Global Surveyor. The IMP was able to obtain full multi-spectral panoramas of the surface and underlying materials exposed by the rover.

Magnetic Properties and Soil Mechanics of the Surface. Magnetic targets were distributed at various points around the spacecraft. Multi-spectral images of these targets identified the magnetic minerals that make up the airborne dust. Using the IMP images, it was possible to identify the mineral composition of the rocks. Detailed examination of the wheel-track images also gave a better understanding of the mechanics of the soil surrounding the landing site.

Atmospheric Structure as Well as Diurnal and Seasonal Meteorological Variations. The Atmospheric Structure Instrument/Meteorology (ASI/MET) experiment was able to monitor the temperature and density of the atmosphere during Entry, Descent and Landing (EDL). In addition, three-axis accelerometers were used to measure atmospheric pressure during entry. Once on the surface, meteorological measurements such as pressure, temperature, wind speed and atmospheric opacity were obtained on a daily basis. Thermocouples mounted on a one meter (3.3 foot) high mast examined temperature profile with height. Wind direction and speed were measured by a wind sensor mounted at the top of the mast, as well as three windsocks interspersed at different heights on the mast. Understanding this data was important for identifying the forces that act on small particles carried by the wind. Regular sky and solar spectral observations using the IMP monitored windborne particle size, particle shape, distribution with altitude and the abundance of water vapor.

Rotational and Orbital Dynamics of Mars. The Deep Space Network (DSN), by using two-way X-Band and Doppler tracking of the Mars *Pathfinder* lander once it was on the surface, was able to address a variety of orbital and rotational dynamics questions. Spacecraft ranging involves sending a code to the lander and measuring the time required for the lander to echo the code back to the Earth-based station. By dividing this time by the speed of light, results can be accurate within 1 to 5 meters (3 to 16 feet) of the distance from the station to the spacecraft. As the lander moves relative to the tracking station, the velocity between the spacecraft and Earth causes a Doppler shift in frequency. Measuring this frequency shift provided an accurate measurement of the distance from the station to the lander. After a few months of observing these features, the Mars *Pathfinder* lander location was determined within a few meters. Once the exact location of *Pathfinder* had been identified, the orientation and precession rate of the pole can be calculated and compared to measurements made with the *Viking* landers 20 years ago. Measurement of the precession rate allowed direct calculation for the moment of inertia. Measurements similar to these are used on earth to determine the makeup of the earth's interior.

Surface Science Phase. After receiving data indicating the health of the spacecraft and a successful landing, commands were sent to the spacecraft to unlatch the IMP camera and the high gain antenna. The first task of the lander was to determine the location of the Sun. To do this, the IMP scanned the horizon for the brightest spot on the horizon. Once the Sun's location was determined, the high gain antenna was directed towards Earth and the first images were received around 4:30 p.m. PDT. The first received data included the mission success panorama, stereo images of both rover ramps, and spacecraft engineering data which included the health status of the spacecraft and the status of the airbag retraction. See Fig. 14.

After examining the imagery, it was determined that the airbags had not fully retracted from the rover petal and that it would not be safe to deploy the rover petals. Commands were sent up to reclose the rover petal, retract the

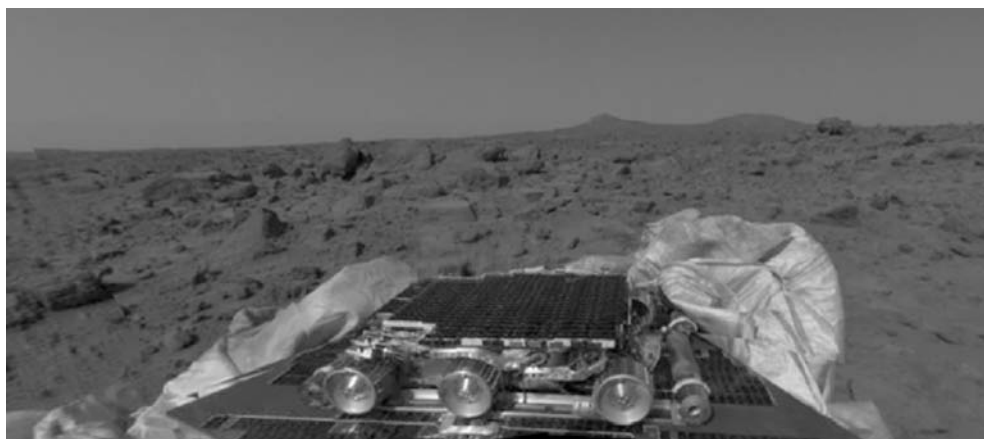


Figure 14. Mission success panorama acquired on July 4, 1997. (Image courtesy of JPL/NASA.)

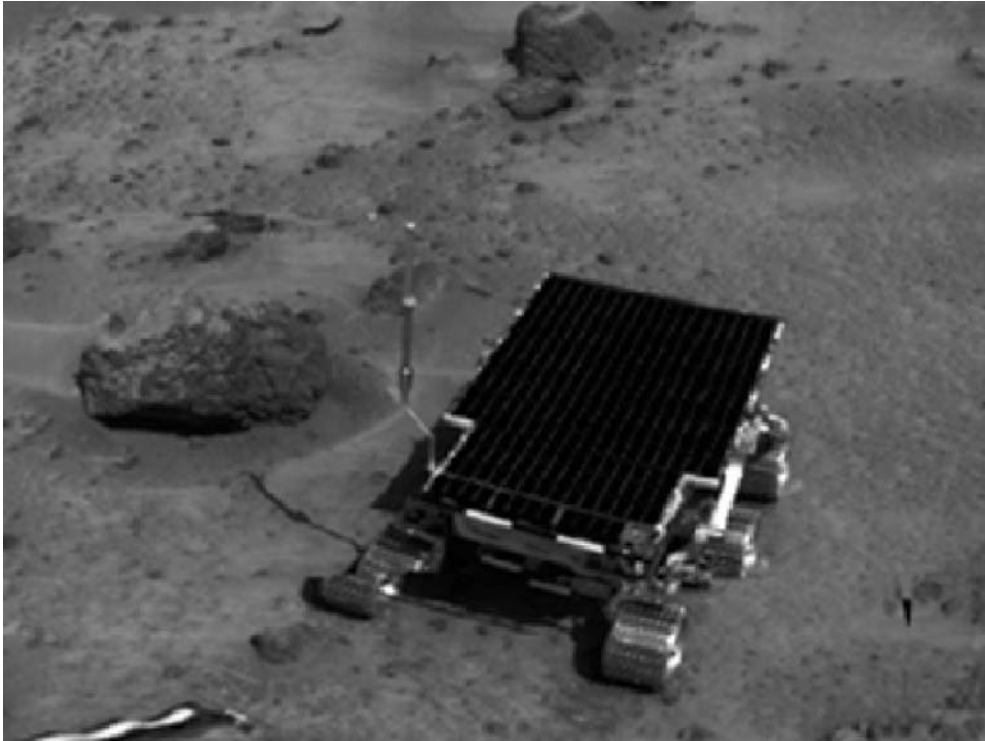


Figure 15. Sojourner on the surface of Mars. Rock to the left is Barnacle Bill. (Image courtesy of JPL/NASA.)

airbag further, and then redeploy the rover petal. After careful examination of a second set of images, the ramps were determined to be safe and the rover was commanded to stand up. A full panorama of the landing site was also returned on the first day of operation and the rover was driven down the rear ramp the following day (Sol 2). After it was determined to be safe to deploy the IMP camera, the camera mast was deployed to a height of 0.8 meters (2.6 feet) at the end of Sol 2. After some minor communication errors between the rover modem and the lander, the rover deployed the APXS at the surface for its first soil sample. See Fig. 15.

Lander Site Location

When the first images had arrived, five prominent horizon features and two small craters were identified in both lander horizon and *Viking Orbiter* images. This enabled the lander to be located within 100 meters (330 feet) of other surface features at 19.13°N, 33.22°W in the U.S. Geological Survey reference frame.

Characteristics of the landing site were determined to be consistent with its prelanding predication of a flat, level flood plain composed primarily of materials left behind by the Tiu and Ares catastrophic floods. The surface is composed of pebbles, cobbles and boulders that closely resemble depositional surfaces found

from catastrophic floods on Earth. Two nearby peaks identified as “Twin Peaks,” appear to be streamlined hills in IMP images; this is consistent with prelanding predictions of *Viking Orbiter* images of the region. Rocks identified in the Rock Garden are imbricated in the direction of the predicted flow; again agreeing with prelanding predictions. Troughs are also visible throughout the scene and have been interpreted to be erosional features produced by the turbulent flood waters. Large boulders can be found perched on top of smaller rocks (i.e. Yogi), consistent with deposition by a flood. Except for later eolian activity, the site appears little altered since it formed up to a few billion years ago.

A variety of soil types have also been identified at the *Pathfinder* landing site. See Fig. 16. These soils appear to be composed of poorly crystalline ferric-bearing materials. Elemental compositions of soil units measured by the APXS are similar in composition to those measured at both of the *Viking* landing sites. Due to the distance between the *Pathfinder* site and the two *Viking* landing sites, the similarities in soil compositions suggest that the compositions are influenced by globally distributed airborne dust.

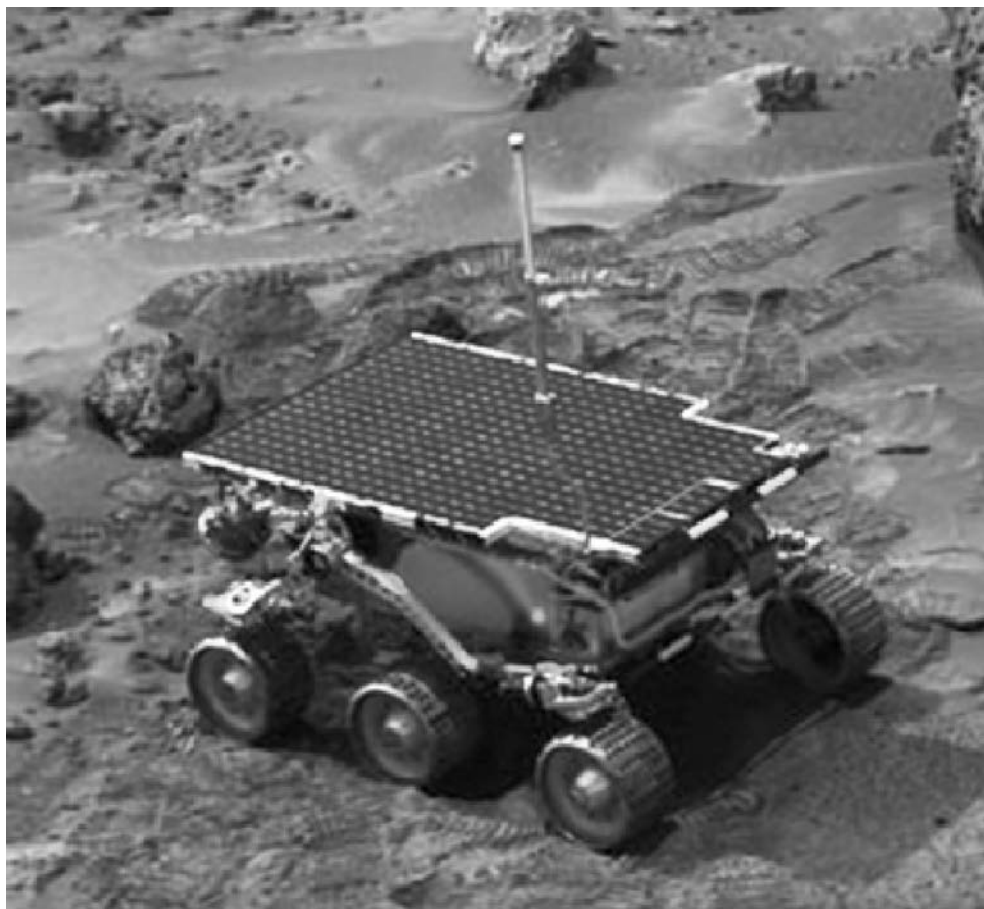


Figure 16. End of day image of the rover. (Image courtesy of JPL/NASA.)

Rocks that have been identified at the *Pathfinder* site are primarily dark gray and partially covered with coatings of bright dust and/or weathered surfaces. See Fig. 17. From the rock chemistry measured by the APXS they appear to be similar to basalt, basaltic andesites, and andesites found on Earth. Rover close-up and IMP images display rocks with a variety of different morphologies, textures and fabrics. Some of the rocks have been hypothesized to be conglomerates composed of rounded pebbles embedded in a finer matrix. Rocks such as these may be the source of numerous rounded pebbles and cobbles that were identified throughout the site. If these rocks are conglomerates, their formation suggests that running water was present to smooth and round the pebbles and cobbles over long periods of time. The rounded materials would then be deposited into a finer grained sand and clay matrix and lithified before being carried to the Ares site. This suggests a warmer and wetter past in which liquid water was stable on the surface.



Figure 17. End of day IMP image of the rover in the Rock Garden. (Image courtesy of JPL/NASA.)

The magnetic properties experiment identified the airborne magnetic dust that was deposited on most of the magnetic targets. The dust is characterized as a light yellowish brown, with clay-sized silicate particles and a small amount of a magnetic mineral (believed to be maghemite). The present interpretation for the maghemite formation is that iron was dissolved out of crystal materials by water and that the maghemite is a freeze-dried secondary precipitate.

Observations from wheel tracks and soil mechanics experiments illustrate that the subsurface consisted of a variety of different materials with different physical properties. See Fig. 18. Rover tracks observed in bright drift material preserved individual cleat marks indicating that they are compressible deposits of very fine-grained dust. Several cloddy deposits found at the landing site appear to be composed of poorly sorted dust, sand-sized particles, lumps of soil, and small rock granules and pebbles.

The atmospheric opacity was determined to be 0.5 and changes slightly higher at night as well as early in the morning due to clouds. The sky is a light



Figure 18. Rover on Mermaid Dune. (*Image courtesy of JPL/NASA.*)

yellowish brown color composed of micron-sized particles and water vapor. See Fig. 19. The upper atmosphere, above 60 kilometers (36 miles) altitude, was determined to be very cold and different from warmer measurements obtained by both *Viking* landers. The differences in the measurements obtained by both *Viking* landers may be attributed to seasonal variations at the time of landing. The multiple peaks in the landed pressure measurements and the entry and descent data are indicative of dust uniformly mixed in a warm lower atmosphere.

The meteorology measurements at the site identified diurnal and higher order temperature fluctuations. The barometric minimum was reached at the site of Sol 20 indicating the maximum extent of the winter south polar cap. Temperatures changed abruptly with time and height in the morning; these observations suggest that the warming of the cold morning air by the Sun created upward moving small eddies. Winds were fairly consistent, and dust devils were detected repeatedly throughout the mission.



Figure 19. Clouds observed at the Ares Valles landing site. (Image courtesy of JPL/NASA.)

Daily Doppler tracking and less frequent two-way ranging during communication sessions between the spacecraft and Deep Space Network antennas resulted in a solution for the location of the lander and the direction of the Mars rotation axis. Combined with earlier results from the *Viking* landers, the estimated precession constant constrains the core radius of Mars to be between 1,300 and 2,000 kilometers (780 to 1200 miles).

From all of the scientific results that have been completed so far, early Mars appears to have been very similar to an early Earth. Some of the materials that make up the crust may be similar to terrestrial continental crust materials. The rounded pebbles, cobbles suggest a possible conglomerate, which supports water rich early Mars. This would imply that the early environment of Mars was warmer and wetter than today and liquid water may have been in equilibrium. Further Mars missions may be able to answer these questions.

R.C. ANDERSON
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PLANETARY EXPLORATION SPACECRAFT DESIGN

Introduction

Having only begun in the early 1960s, planetary-exploration spacecraft design is a fairly new art (1). Between 1962 and 1973, NASA designed, built, and launched 10 spacecraft named Mariner to explore Venus, Mars, and Mercury for the first time. In 1962, the Jet Propulsion Laboratory (JPL) launched the Mariner 2 spacecraft (Fig. 1) to explore Venus (2). In 1965, Mariner 4 flew by Mars and captured the first close-up photos of another planet.

Each year since 1962, NASA, the USSR, Japan, or the European Space Agency has launched at least one planetary-exploration spacecraft (3). Two of the best known are NASA's Voyagers (4–6) (Fig. 2), which were launched in 1977, two years after NASA's Viking mission (7) sent landers and orbiters to Mars (Fig. 3). Today, 24 years after launch, Voyager 2 has completed a tour of Jupiter, Saturn, Uranus, and Neptune, and still returns data from a distance of 64 AU (astronomical units, the average Earth–Sun distance; one AU is 150 million kilometers), which takes nearly nine hours of light time. NASA spacecraft have visited every planet in the solar system except Pluto, and orbited Venus, Mars, and Jupiter. In 1975, the USSR's Venera 9 and 10 spacecraft sent the first photos from the surface of another planet (Venus); in 1986, the European Space Agency's Giotto spacecraft captured the first detailed photos of a comet's nucleus (Halley's). The NASA/JPL Cassini spacecraft is due to enter Saturn orbit in 2004 (8). The NASA/APL MESSENGER spacecraft is due to orbit Mercury in 2009 (9). MESSENGER is being built by the Johns Hopkins Applied Physics Laboratory (APL),



Figure 1. Mariner 2, launched in 1962, became the first spacecraft to fly by another planet, studying the Venusian atmosphere and surface. During its $3\frac{1}{2}$ -month journey to Earth's neighbor, the craft made the first-ever measurements of the solar wind, a constant stream of charged particles flowing outward from the Sun. It also measured interplanetary dust, which turned out to be scarcer than predicted. In addition, Mariner 2 detected high-energy charged particles coming from the Sun, including several solar flares, as well as cosmic rays from outside the solar system. As it flew by Venus on 14 December 1962, Mariner 2 scanned the planet with infrared and microwave radiometers, revealing that Venus has cool clouds and an extremely hot surface. Mariner 2's signal was tracked until 3 January 1963. The spacecraft remains in orbit around the Sun.

which also built the Near Earth Asteroid Rendezvous 10 (NEAR) spacecraft (Fig. 4) that orbited and touched down on the asteroid Eros in February 2001.

This article addresses the important elements of planetary-exploration spacecraft (PES) design. The scope of this article includes those spacecraft that fly by, orbit, or land on other planets, asteroids, or comets. Planetary-exploration spacecraft are truly fascinating machines, and their missions are some of humankind's most audacious endeavors. Spacecraft elements are described in some detail, and the principal focus is on the system design and the technical subsystems. Ground stations, launch vehicles, science, and navigational elements (11) are not addressed.

Requirements and Constraints on Planetary-Exploration Spacecraft Design

Introduction. PES design is a process that is driven by the unique nature of planetary-exploration missions and their scientific investigations. Mission and science requirements are used to determine spacecraft system-level requirements, and in turn subsystem requirements, as shown in Fig. 5. Basically, where the mission goes, how it gets there, and what it does drives the spacecraft design.

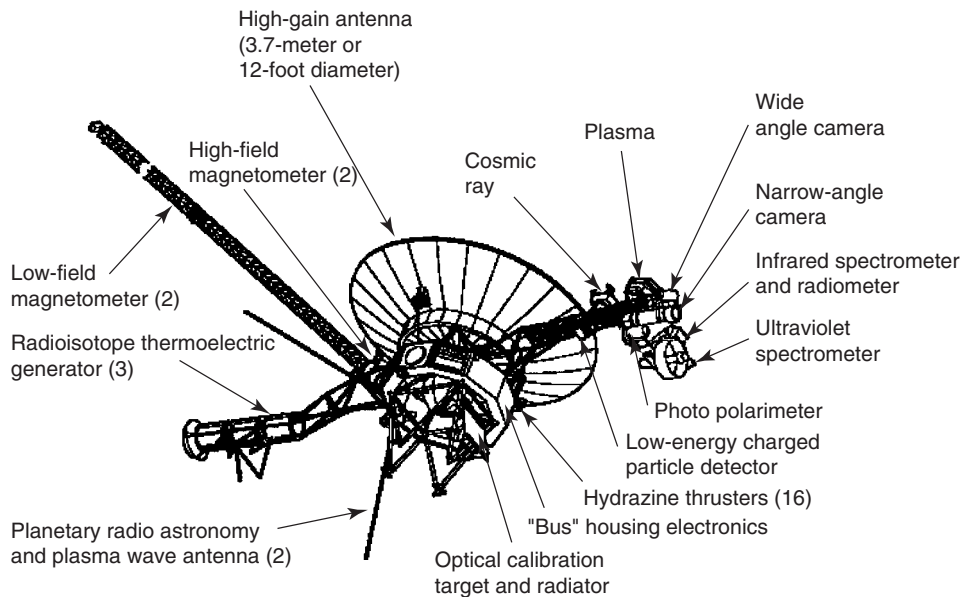


Figure 2. The twin spacecraft Voyager 1 and 2 flew by and observed Jupiter and Saturn; Voyager 2 went on to visit Uranus and Neptune. In 1998, Voyager 1 became the most distant human-made object in space.

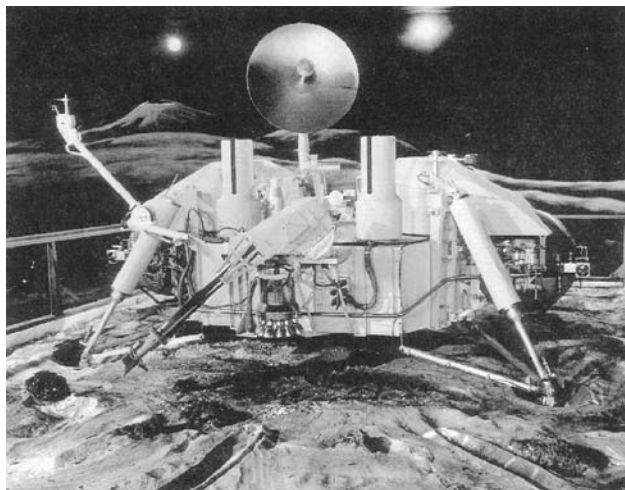


Figure 3. The Viking Mars Lander spacecraft is shown here. Viking 1 landed on Mars on 20 July 1976, and Viking 2 landed on 3 September 1976. Viking 1 was the first successful soft landing of a spacecraft on Mars. The two Viking spacecraft were identical, and they sent back photographs of the Martian surface. They also collected a wealth of data on atmospheric composition, meteorological conditions, soil samples, and seismic activity. Each Viking spacecraft also carried a set of experiments that looked for biological activity on the surface of Mars. The results were negative. Viking 1 transmitted data for more than six years. Viking 2 was shut down on 11 April 1980.

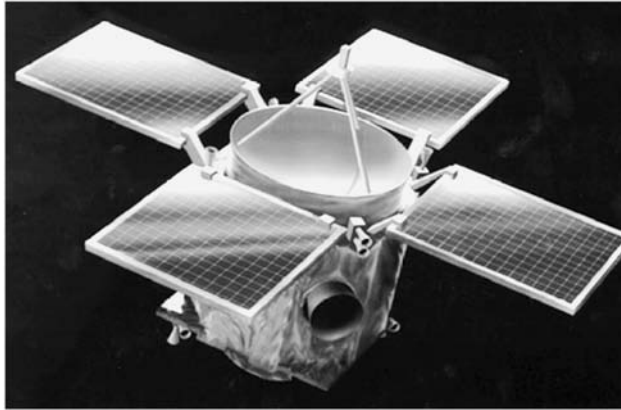


Figure 4. In 1996, NEAR-Shoemaker was the first Discovery Program spacecraft to be launched, and five years later, in February 2001, it became the first ever to orbit and land on an asteroid. Using six highly specialized instruments to gather data about its primary target, asteroid 433 Eros, NEAR has answered many fundamental questions about the nature and origin of asteroids and comets. The spacecraft snapped 69 detailed pictures during the final 3 miles (5 km) of its descent, the highest resolution images ever of an asteroid that showed features as small as one centimeter across.

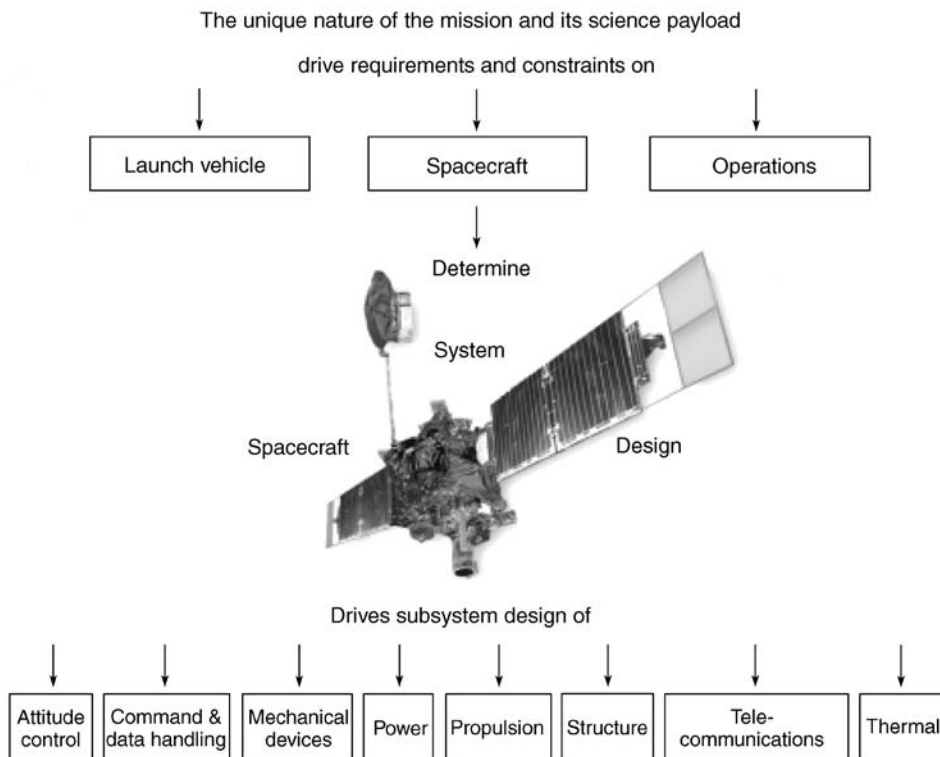


Figure 5. Planetary spacecraft design flow.

The remainder of this section presents a brief discussion of important mission and system requirements and constraints that drive the design of a planetary-exploration spacecraft.

Mission Objective. The mission objective is to perform scientific exploration and investigation. To conduct the desired scientific investigation, it is necessary to transport the scientific payload to the destination and provide power, environmental protection, the proper instructions and orientation for performing measurements and the necessary communications technology for transmitting the resulting data to Earth. A large complex of people, computers, communication lines, and tracking antennae on Earth is needed to guide and support the mission.

Very High Standards. Two paramount constraints for all missions are that, once launched from Earth, spacecraft hardware can never be repaired, and there is very low tolerance for any failure. Because overwhelming media coverage ensures that the world will instantly learn of any problem, no space agency that commits public resources to build, launch, and operate a PES welcomes even the perception of failure. All of this creates enormous pressure on a PES and its mission to succeed. Consequently, the spacecraft must be designed, built, and operated to extraordinarily high standards.

Instruments. The fundamental function of a PES, (what it does) is to transport the instruments needed to perform the scientific and exploratory investigations. Most of these instruments are passive devices that operate in parts of the electromagnetic spectrum, including long-wavelength radio waves, the infrared (IR), visible, the ultraviolet, X rays, and gamma rays. Some instruments, such as radar and light detection and ranging (LIDAR), are active-sensing. The instruments place requirements on the spacecraft, such as mass, power, pointing, data rate, and data volume. Some instruments place special requirements on the PES. Spacecraft that carry sensitive magnetometers must be “magnetically clean”; that is, designs must specify special electronic parts, shielded components, and sometimes equipment to generate and maintain offsetting magnetic fields. To avoid receiving its own spacecraft emissions, for example, a magnetometer may need to be flown at the end of a long boom; this greatly affects the attitude control-system design. A gamma-ray detector onboard means that overall spacecraft emissions must be checked and components altered or removed so that their output does not swamp the target planet’s signals.

Planetary Protection. Planetary protection is a unique requirement for planetary spacecraft. It means protecting another planet’s environment from Earth’s biological contamination and protecting Earth’s environment from extraterrestrial contamination. Usually applied to landers, it is critically important for those missions whose objective is to return a sample of another planet to Earth; it also applies to orbiters whose orbit may someday degrade, causing the spacecraft to hit the surface. To illustrate, consider the Mars orbiter that NASA currently plans to launch in 2005. The planetary protection requirements for this mission are that

- all flight hardware is to be assembled and maintained in class 100,000 clean rooms (or better);
- the probability of an impact on Mars by any part of the launch vehicle (including the upper stage) that leaves the vicinity of Earth must not exceed 10^{-4} ; and

- the probability of an impact on Mars by the orbiter due to all causes must not exceed 0.01 for the first 20 years after launch and 0.05 for the succeeding 30-year period. To meet part of the third requirement, the flight system must be designed with a reliability for the mission (launch through the attainment of the final orbit) better than 0.99, and the spacecraft-ballistic coefficient must be considered.

Solar System. The most important influence on the design of a planetary-exploration spacecraft is the solar system itself. A spacecraft designed to travel the solar system must successfully operate in the solid-particle and radiative environments of the solar system. The spacecraft must operate far from Earth and at a Sun range greatly different from 1 AU. The large Earth range implies a significant light time for signals sent to and from the spacecraft, long trip times to the final destination, and the need for precise navigation. Launch dates for PES are inflexible, may occur only once every 2 years (or less frequently), and are set by the well-known, but totally inflexible motion of the planets around the Sun.

Rigid and Infrequent Launch Windows. Solar system dynamics are well known, totally fixed, and make for rigid, inflexible, and infrequent launch windows separated by many years—even decades. There is enormous pressure to deliver interplanetary missions to the launch pad on time because missing a window results in intolerable consequences for missions and careers alike. For Mars, for example, a mission may have to wait two years for the next opportunity. Missions to the outer planets may wait much longer. The alignment of the planets Jupiter, Saturn, Uranus and Neptune that the Voyager 2 spacecraft used for minimum trip time occurs about once every 175 years. In contrast, if an Earth-orbiting spacecraft arrives late for a launch window, another one will open in the following week or month. Compared to a typical Earth orbiter, a planetary spacecraft's launch opportunities are very infrequent, and the consequences of failure are especially severe. This consequence of failure creates a greater need for high reliability that is usually implemented through greater robustness in design (parts quality, fault protection, and use of redundancy).

Round-Trip Light Time. An important spacecraft-design constraint is round-trip light time—the minutes or hours that it takes for a distant spacecraft's signal to reach Earth, plus the control-signal return time. Because of the solar system's vast expanse, a typical Earth-to-PES distance is usually many orders of magnitude greater than the Earth-to-Earth-orbiter distance. A PES at 1 AU from Earth, for example, is 4200 times farther away than a geosynchronous Earth orbiter (GEO), which receives a ground signal from Earth in only a tenth of a second. A spacecraft at Mars may be as much as 2.5 AU from Earth—a one-way, light-travel time of 20 minutes. At Saturn, Cassini's maximum distance to Earth requires a one-way, light-travel time of more than 1.5 hours. Despite PES radio signals traveling at the speed of light, the spacecraft designer must consider the long time signals take between spacecraft and Earth.

To minimize mission risk so that it is similar to that of an Earth orbiter, distant spacecraft must operate more autonomously, especially regarding fault detection and recovery. Long round-trip light time affects the way planetary spacecraft are commanded. An Earth orbiter can wait for a command signal before it transmits its recorded telemetry; a planetary spacecraft cannot.

Spacecraft that perform split-second events, such as landing on Mars, cannot afford to wait for a command to be acknowledged before another is sent. For example, if a spacecraft transmits a distress signal, while executing split-second events, such as entering Mars' atmosphere, Earth would receive it only after the probe is on Mars' surface—too late for ground controllers to help.

Earth Range. A spacecraft's telecommunications subsystem is especially driven by a planetary explorer's large Earth range (distance to Earth). The strength of the received signal is inversely proportional to the square of the range. The power per unit area received from a GEO spacecraft is 4200 squared or 72 dB larger than that from the planetary spacecraft at a range of 1 AU from Earth. To receive low-strength signals, NASA built special ground stations collectively called the Deep Space Network (DSN) (12) (Fig. 6). The DSN is an international asset that tracks planetary spacecraft from many countries, and the DSN receiving network has special interface requirements for planetary spacecraft. Because navigating interplanetary space requires extremely accurate clocks, the DSN uses a hydrogen-maser-based frequency reference whose accuracy is equal to the gain or loss of 1 second in 30 million years. The maser is a reference that generates an extremely stable uplink frequency that the spacecraft uses, in turn, to generate its coherent downlink.

Even using the DSN, a planetary probe's uplink and downlink data rates are lower than those of an Earth orbiter. If a planetary spacecraft is turning very quickly because of a mission need or tumbling because of a fault, low uplink rates can preclude commanding. Because of the high signal strength, an Earth orbiter



Figure 6. The Deep Space Network (DSN) is an international network of antennae that supports interplanetary spacecraft missions and radio and radar astronomy observations that explore the solar system and the Universe. Via its vital, two-way communications link, the network guides and controls interplanetary explorers and brings back the images and new scientific data that they collect. All DSN antennae are steerable, high-gain, parabolic-reflector antennae.

can receive a command in nearly any attitude, but distant planetary probes must be in specific attitudes or they will not receive the signal. A limited downlink rate can strongly drive data storage, data compression, and optimal recording of data. A limited downlink rate may also mean that a spacecraft can transmit only a small part of the data that its instruments can produce.

Mission Energy and Trip Time. The destinations for planetary spacecraft are energetically a “long way from Earth.” The mission energy manifests itself not only in the required change in spacecraft velocity (known as ΔV) for departure from Earth and after launch but also in trip time to the destination. Because the required mission energies are so large, most planetary mission trajectories are “minimum energy” trajectories (13). A consequence of these “minimum energy” trajectories and the vast distances are long trip times to the destination. The largest ΔV for Earth-orbiting spacecraft is normally required by a geosynchronous communication spacecraft where the ΔV required to go from low Earth orbit (LEO) to GEO is about 4200 m/s. The total ΔV beyond LEO for some typical planetary missions are as follows. For a low, circular Mercury orbiter where the trip time is about 4 years, the required ΔV is 7000 m/s. A direct 0.4-year trip to a low Venus orbit requires about 6900 m/s. For a Pluto orbiter that has a 20-year trip time, 15,000 m/s is required. Usually, the launch vehicle for both Earth orbiters and planetary spacecraft supplies part of this ΔV . Several of these planetary missions have not been performed because the combination of ΔV and trip time is so large.

Because of the large ΔV for many planetary missions, methods besides traditional chemical propulsion have been developed, such as electric propulsion systems and the use of the target-body atmosphere to aid in capture at the planet. Aerobraking (14,15), as this is known, is a technique wherein the spacecraft is deliberately flown into the top of the target planet’s atmosphere. There, the drag on the spacecraft acts like a brake, and spacecraft energy is dissipated. Aerobraking uses almost no propellant to provide ΔV to change the kinetic energy of the spacecraft. The NASA/JPL Mars Global Surveyor (16) spacecraft (Fig. 7) and also the Mars Odyssey (17) spacecraft (Fig. 8) that arrived at Mars in 2001 used aerobraking.

Long trip times and minimum-energy trajectories require that planetary spacecraft be extraordinarily reliable. Some Earth orbiters may live as long as 15 years, but their operations in orbit begin only a few weeks after launch. For many planetary spacecraft, their missions do not begin until they reach their targets, which could take years. For example, Voyager 2 traveled 12 years before it successfully encountered Neptune. The Cassini probe must travel for 6.6 years before its Saturn orbital operations begin. So, it is many years after launch that these spacecraft maneuver for orbit insertion and open into their final mechanical configuration. The ship’s mechanisms, components, and instruments must remain workable during those years traveling through space and when activated, must perform perfectly.

Extended missions also affect spacecraft operations. For example, when cruise time is lengthy, additional operational modes, software, and procedures are needed. Because of personnel turnover during missions measured in decades, retraining is also absolutely critical if fatal errors are to be avoided in the mission’s end game, when the people who designed the final procedures have moved

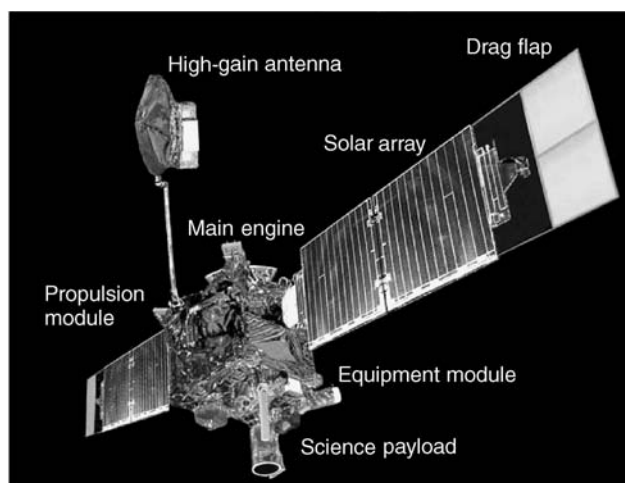


Figure 7. The Mars Global Surveyor (MGS) entered Mars' orbit in late 1997 and was the first successful mission to the red planet in two decades. MGS scientific instruments include a high-resolution camera, a thermal-emission spectrometer, a laser altimeter, a magnetometer/electron reflectometer, an ultrastable oscillator, and a radio-relay system. Beginning its prime mapping mission in early 1999, MGS studied the entire Martian surface, atmosphere, and interior from a low-altitude, nearly polar orbit during one Martian year (almost two Earth years). Having completed its primary mission in early 2001, it continues in an extended mission phase. MGS has returned more data about the red planet than all other Mars missions combined.

on. To reduce peak-year costs for multiyear-missions, command sequences for planetary encounters are not developed until after launch. Therefore, spacecraft design must accommodate new flight software that will be tested “on the fly.” Changes in data-system technology also affect lengthy missions, and vice versa. Ground-system components, for example, may be upgraded during the mission. On the positive side, for long-cruise-time missions, there is time to diagnose and hopefully correct any errors before the primary mission begins.

Space Radiation Environments. Because the planetary-spacecraft environment is very different from that of an Earth orbiter, it also impacts the system. Although usually stable, this distant environment is uncertain; one of the mission goals may be to learn more about it. To keep the risk level close to that of an Earth orbiter, a distant mission may require more environmental analysis and margin. PES design must accommodate space radiation and solid particles. The three types of space-radiation environments are planetary-trapped radiation, galactic cosmic radiation, and solar-energetic particles.

Because the trip through them is short, spacecraft are not affected by Earth's radiation belts, but other planetary belts can be a serious threat. Even short gravity-assist passages through Jupiter's system, for example, will expose the spacecraft to most of a mission's total radiation dose. Heavy-ion fluxes will also disrupt critical mission functions that have single-event effects (SEE).

Galactic cosmic radiation (GCR) is composed of high-energy nuclei. Earth's magnetic field provides some shielding against this radiation for spacecraft in low-to-medium Earth orbits. Even so, the radiation flux from GCR in

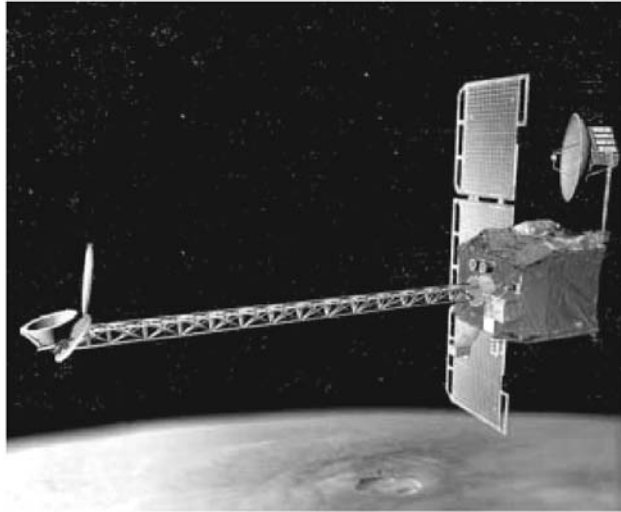


Figure 8. Mars Odyssey has a mass of 758kg and carries scientific instruments, including a thermal emission imaging system, a gamma-ray spectrometer, and a Mars radiative environment experiment. It arrived at Mars in late 2001. The Odyssey orbiter's mission is to determine the planet's surface composition; detect water and shallow, buried ice; and to study the radiation environment. In May, 2002 Odyssey scientists announced that their instruments had detected enormous quantities of water ice just below Mars' surface. It has long been thought that Mars' surface consists of rock, soil, and icy material, but the exact composition is largely unknown. Odyssey will help identify soil minerals and surface rocks, and it will study small-scale geologic processes, as well as future landing-site characteristics. By measuring the amount of hydrogen in the entire planet's upper meter of soil, the spacecraft will uncover the amount of water available for future exploration and also more clues about Mars' climatic history. The orbiter will also collect radiation data to help assess risks to future human explorers, and it will act as a communications relay for future Mars Landers.

interplanetary space is low and does not contribute significantly to the mission's total ionizing dose, but sufficient heavy ions are present in the GCR populations to create a SEE threat to some electronic parts throughout the mission.

Solar-energetic particles are also attenuated by Earth's magnetic field, but in interplanetary space, fluxes due to a given solar event may be two to three times more intense than near Earth and thus present both a total dose and SEE threat. When they are outside Earth orbit, solar-event fluences dissipate in inverse proportion to the square of the heliocentric distance. So, outer planet missions are reasonably safe. But inside 1 AU, it is a different story. The hazards of solar-event radiation can be considerable for missions that visit Venus or Mercury. For instance, even a moderate solar event can incapacitate a spacecraft that passes close to the Sun.

The solid-particle danger to interplanetary spacecraft has two origins: micrometeoroids (particles that orbit the Sun) and dust (particles that orbit other solar-system bodies). Artificial debris is encountered only in a mission's immediate post launch phase. Although there are few particles in interplanetary space, encounter velocities are high ($\sim 15\text{--}50\text{ km/s}$) and can be very damaging. The nature of the encounter dictates the velocity. Because Cassini will go into a

Saturn orbit, its encounter velocity with the planet's ring particles is much lower than Voyager's speed when it passed close to Saturn to obtain a gravity assist.

Sun Range. The Sun range (distance to the Sun) especially affects PES power and thermal-control subsystems. Widely varying spacecraft–Sun distances—especially when they change during a mission—dramatically impact the design. For example, Cassini flew by Venus (0.6 AU) for gravity assists and headed out to Saturn (10 AU) to conduct its prime mission. A Mercury orbiter (18) encounters a huge thermal range: its illuminated side receives 10.6 suns (mainly short-wave IR), and the other side sees 8.5 suns (long-wave IR reflected from the planet's surface). And a few hours later in solar eclipse, the orbiter will see no “suns” on either side.

Sun-range effects on the power subsystem are directly opposite, depending on whether the mission destination is less than or greater than 1 AU. Spacecraft going to the inner planets have the problem that their power output increases with mission time, up to a point where solar cells developed for Earth orbiters are no longer usable without special coatings, or they are not usable at all. The spacecraft may solve its excess-heat problem by pointing away from the Sun or by shunting the excess power and radiating the excess as heat.

The opposite problem is true for a spacecraft bound for an outer planet, because its solar-array power gradually drops off and finally stops. Special solar arrays designed for low-light, low-temperature conditions are required. In some cases, a PES will use a radioisotope thermoelectric generator (RTG) for electric power production. Another problem for outbound spacecraft is that unless “makeup heat” is provided or heat is conserved, components may cool below safe operating temperatures. Experience dictates that outer-planet missions must invest more of their development time and money in power-efficient technologies.

Navigation (11). Because of the size of the solar system and the motions of the planets, a planetary-exploration spacecraft must be carefully navigated to its destination. Radiometric data can be used to determine spacecraft position relative to a planet—to within a few kilometers at best. Velocity data are accurate to within a few tenths of a mm/s. Nongravitational acceleration caused by unplanned or unmodeled spacecraft accelerations such as outgassing, leakage, or uncoupled attitude-control maneuvers can greatly affect spacecraft navigation. Acceleration due to these forces must be kept within a few mm/s², accurately modeled, and reported. In 1999, perhaps because of an error in modeling these forces, engineers believe that the Mars Climate Orbiter (19) entered the Mars atmosphere too low to achieve orbit and probably burned up.

Planetary-Exploration Spacecraft Subsystem Design

Subsystems are individually tailored and integrated to implement the overall system design (20). Subsystems include structure, thermal, mechanical devices, command and data handling, attitude control, telecommunications, power and propulsion. Table 1 presents the subsystem characteristics of several planetary-exploration spacecraft.

Structure. The structural subsystem includes primary and secondary mechanical and structural members that support and align all spacecraft flight

Table 1. **System Characteristics of some NASA Spacecraft**

Project & launch year	Engineering subsystem mass, kg	Science instrument mass, kg	Total dry mass, kg	Propellant mass, kg	Max earth range, AU	Max downlink, kb/s, (receiver antenna size, m)	Total Δ , m/s	Pointing, mrad (3 sigma)	Data storage, Mb	Beginning of life, power, W
Mariner 4 1964	235	16	251	10	2.3	0.033 (26 m)	60		5	700
Mariner 6 & 7 1969	355	58	413	10	2.5	0.033 (26 m) 16.2 (64 m)	59		23	830
Mariner 9 1971	438	68	506	491	2.5	16.2 (64 m)	1650	25	180	830
Pioneer 10 & 11 1972, 1973	199	33	232	27	70 +	0.256 (26 m) 2.048 (64 m)	200	15	49	155
Mariner 10 1974	396	78	474	29	1.7	117.6 (64 m)	119		180	490
Viking 1 & 2 1975	825	74	898	1440	2.5	16.0 (64 m)	1551	2.5	1120	1400
Voyager 1 & 2 1977	600	117	717	106	40 +	115.2 at 5 AU; 45 at 10 AU (70 m)	380	2.5	500	475
Pioneer Venus Orbiter 1978	476	45	521	32	1.7	0.170 (26 m)	1200	8	1048	226
Pioneer Venus Multiprobe 1978	255	316 + 272	843	32	1.7	4.096 (64 m) 0.170 (26 m) 4.096 (64 m)	150	8	1048	214
Magellan 1989	899	132	1031	128	1.7	276 (70 m)	2885	2.7	3600	810
Mars Observer 1992	931	141	1072	1383	2.5	74 (34 m)	2306	8.1	2100	900
Galileo 1989	1052	105	1165	957	5.2	134 (70 m)	1650	2.5	900	570
Mars Global Surveyor 1997	582	77	660	381	2.5	85.3 (34 m)	1290	8.1	3000	605
Cassini 1997	1754	363	2117	3132	10	249 (70 m)	2360	1	3600	800
Pathfinder, Mars Lander 1996	740	9	749	85	2.5	11.06 (70 m)	131	19	1024	270

subsystems and equipment during ground handling, launch, and flight. Structure also protects against natural and induced environments such as vibration, meteoroids, radiation, and electromagnetic interference. Generally, the structure subsystem of an Earth-orbiting and a planetary spacecraft are alike.

Thermal. The thermal subsystem heats or cools spacecraft components to maintain them within operating (and nonoperating) temperature ranges. This function includes meeting special temperature requirements of components such as batteries and cryogenic instruments. Temperature control must be effective throughout the mission's prelaunch, launch, cruise, and operational phases.

Stated simply, this is a planetary spacecraft's thermal-design strategy: Use passive components whenever possible because they use less power, operate more simply, and are thus more reliable than active ones. The idea is to minimize temperature sensitivity due to external and internal heat fluctuations and emphasize passive techniques (for example, blankets, radiators, coatings) over active ones (for example, heat pipes, fluid loops, thermostatically controlled heaters).

Earth orbiters have "in Sun" and "in eclipse" design points, whereas planetary orbiters have "in Sun" and "in eclipse" conditions that may vary dramatically during the mission because of the change in Sun range. To design the thermal subsystem confidently for a planetary mission, additional thermal analyses and testing are required. In addition, for those missions going closer than 1 AU to the Sun, available thermal-control materials and paints may be limited, and entirely new ones may need to be developed. A requirement to keep large amounts of propellant at a given temperature during a long cruise between launch and orbit insertion may be the main power and thermal design driver.

All spacecraft use multilayer blankets for thermal control. For PES, the multilayer thermal-insulation blankets also provide some protection against micrometeoroid impacts. They are made of Kapton, Kevlar, or other fabrics strong enough to absorb energy from high-velocity micrometeoroids before they damage spacecraft components. Impact hazards are greatest when crossing the ring planes of the Jovian planets. Voyager recorded thousands of hits in these regions, fortunately from particles no larger than smoke particles. Spacecraft, sent to comets such as Stardust, (21), carry massive shields to protect them from hits by larger particles.

Mechanical Devices. Mechanical devices perform functions such as separating a spacecraft from its launch vehicle, deploying booms, articulating high-gain antennae and solar arrays, jettisoning aeroshells and deploying parachutes, releasing instrument covers, and controlling fluid flow in propulsion and pressurization systems. These devices must operate at 100% reliability for many years after launch. Earth orbiters achieve their final spacecraft mechanical configuration early in their lifetime. But, because cruise time to a mission site is very long, it may be many months or years before a PES assumes its final mission configuration. The final deployment of the Mars Observer spacecraft's (22) solar panels, high-gain-antenna boom, and a science-instrument boom were to have occurred after insertion at Mars, nearly 1 year after launch.

The Cassini (23) spacecraft will release a large Titan atmospheric probe 7 years after launch. There is always concern that mechanical-actuation devices may be less reliable after a long time in space. Because these devices are candidates for single-point failures, a design remedy is to add redundant actuators.

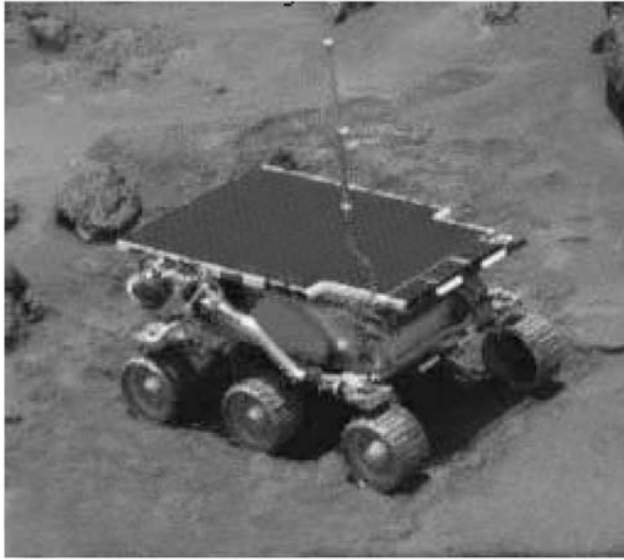


Figure 9. The Mars Pathfinder's Sojourner rover is 280 mm high, 630 mm long, and 480 mm wide. Using rover and lander images, an Earth-based operator controlled the rover. But, because of the light-time delay of 10–15 minutes, an onboard, hazard-avoidance system allowed it some autonomous control as it explored the surface, so it would not always have to wait for commands from Earth. On the rover, 0.2 square meters of solar cells, provide energy for several hours of operations per sol (1 Martian day = 24.6 Earth hours).

Mars Pathfinder, which successfully deployed Sojourner rover (Fig. 9) on Mars, depended on 42 pyrotechnic events during its successful atmospheric entry, descent, and landing. Design practices emphasize large strength or force margins, redundancy where practical (for example, bearings within bearings), vacuum-compatible materials and lubricants, and special attention to the effects of thermal expansion and contraction. All mechanical devices are extensively tested to verify margins and lifetime, and intentional “fouling” tests are conducted to achieve extra confidence. A mechanical actuation system that successfully operated on an Earth orbiter may fail to work on a planetary spacecraft. An example is the Galileo (Fig. 10) spacecraft's high-gain antenna that failed to open completely. Though the actual cause will never be known, extensive prelaunch handling—rather than the long-storage time en route to Jupiter—may have been the culprit.

Command and Data Handling. The brain and memory of a spacecraft are the command and data-handling subsystem (C&DH). Consisting mainly of computers, memory, and input/output devices, it also provides the spacecraft's time reference. The C&DH also performs real-time and preprogrammed functions such as decoding ground commands; distributing discrete and coded commands to subsystems; responding to onboard-generated events; and collecting, managing, and storing scientific and engineering data. Primary design considerations include processor speed, quantity, and access speed of memory, as well as fault-detection-and-correction techniques. In some cases, onboard processing

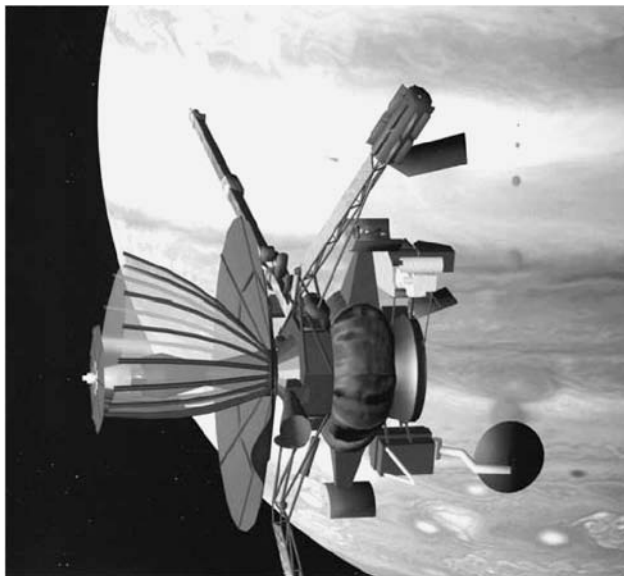


Figure 10. The Galileo spacecraft arrived at Jupiter in late 1995, entered orbit, and dropped its instrumented probe into the giant planet's atmosphere. Since then, it has made dozens of orbits around Jupiter and flown close to each of its four major moons. Galileo discovered that the cratered surface of Jupiter's moon, Europa, is mostly water ice, and there is strong evidence that it may cover an ocean of water or slushy ice. Galileo also found indications that two other moons, Ganymede and Callisto, have layers of liquid saltwater as well. Half of the spacecraft contains pointable instruments, such as cameras, and is held fixed in relation to space; the other half, containing instruments that measure magnetic fields and charged particles, slowly rotates to optimize the measurements. Other major scientific results include details of varied and extensive volcanic processes on the moon Io, measurements of conditions within Jupiter's atmosphere, and the discovery of a magnetic field generated by Ganymede. Despite heavy radiation doses, Galileo continues to return scientific data in an extended mission phase.

of optical-navigation images may be required to reduce the quantity of data transmitted to Earth or to reduce response time for critical events. The need for autonomous fault-diagnosis and recovery on planetary spacecraft requires that designers develop, design, and validate additional onboard software. And in turn, fault-protection software needs computer memory, processing power, and additional telemetry data.

Compared with an Earth orbiter, a planetary spacecraft must have additional intelligence and autonomy to monitor and control itself. The PES is always a great distance from home, is tracked as little as once per day or week, and cannot communicate with Earth during certain periods. For example, during superior conjunctions (when the spacecraft and Earth are on opposite sides of the Sun), autonomous fault-protection software must intervene in case of an onboard failure. Light time and tightly constrained tracking schedules prohibit ground teams, who control and monitor the spacecraft, from immediately responding to onboard anomalies. When failures occur, C&DH fault-protection algorithms must detect them and in the case of a communications interrupt, reestablish contact with Earth. A spacecraft may have many different fault-protection monitoring

algorithms running simultaneously and must be prepared to request C&DH to take action. The command-loss timer, for example, is reset to a predetermined value, for example 1 week, each time a command arrives from Earth. If the timer runs down to zero, it is assumed that the receiver or another component in the command string suffered a failure. The fault-protection response may then be to switch to redundant hardware to reestablish radio contact with Earth.

Other fault-protection responses include requesting safe mode; shutting down or reconfiguring components to prevent damage; or performing an automated, methodical search to re-establish Earth-pointing to regain communications. Though entering safe mode may temporarily disrupt scientific data gathering, it provides reliable spacecraft and mission protection. Read-only memory (ROM) usually carries a minimal set of safe-mode instructions (the Magellan spacecraft (Fig. 11) ROM contained only 1 kB of this code); commands can hide in ROM for the worst imaginable scenarios of runaway-program executions or power outage. More intricate safe-mode and fault-protection routines (for example, “contingency modes”) and parameters for use by the ROM code typically reside in random access memory (RAM), where they can be updated as needed during the mission lifetime.

Attitude Control. The attitude control subsystem keeps spacecraft motions stable and precisely points and orients the spacecraft (or single instruments) despite gravity gradients, propulsion-system torque, solar pressure, and sometimes micrometeoroid impacts. Precise pointing is important because the high-gain antenna has to acquire Earth contact; scientific instruments must collect data; thermal radiators have to be properly positioned; Sun-and-star sensors

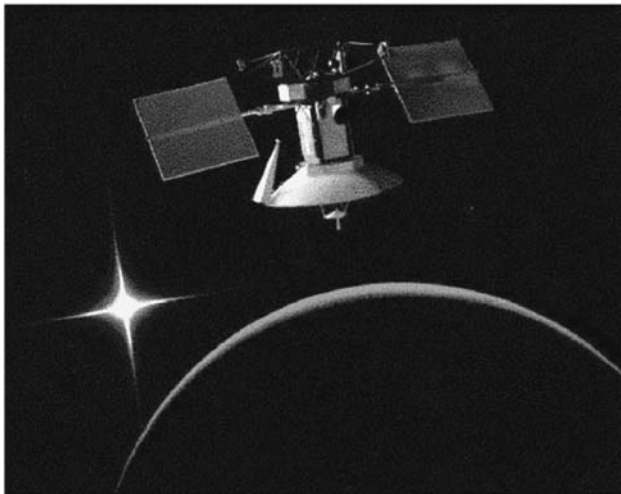


Figure 11. The Magellan spacecraft was carried into Earth orbit by the Space Shuttle Atlantis and was propelled to Venus by a solid-fuel motor. Entering orbit in late 1990, Magellan pierced Venus’ veil of swirling clouds, mapped its surface with its imaging radar, and dramatically improved upon the mapping resolution of orbiters that the United States and the Union of Soviet Social of Republics sent in the 1970s and early 1980s. Flight controllers tested a new maneuvering technique called aerobraking that uses a planet’s atmosphere to slow a spacecraft. At mission end in 1994, Magellan’s orbit was lowered a final time, and it plunged to the planet’s surface; contact was lost the following day.

must provide attitude references; and to generate maximum power, solar panels have to be Sun-oriented. The spacecraft must also point in the correct direction to make precise trajectory-control maneuvers.

Because large amounts of propellant can be expended, the attitude-control subsystem design must take into account the resulting dramatic changes in a spacecraft's inertial properties as a function of mission time. A PES is not orbiting Earth, so the attitude-control subsystem must operate differently. Spacecraft orbiting other planets sometimes use horizon sensors, so Earth-horizon sensors must be adapted to a different atmosphere or no atmosphere. GPS receivers cannot be used for orbit or attitude information. Without Earth as an attitude reference, planetary spacecraft become highly dependent on celestial sensors and celestial-attitude determination. Without a well-characterized, adequately strong magnetic field, magnetic-torque rods cannot be used for attitude maneuvers or to desaturate reaction wheels. Planetary spacecraft may be sent to orbit planets that have magnetic fields (Jupiter), but usually the magnetic field is not well enough known to be used for attitude determination and control.

Because of the various forces acting on it, an unstabilized spacecraft will acquire rotational motion and tumble chaotically. Like Earth-orbiting spacecraft, a PES may be spin-stabilized or three-axis controlled. Spinners that are relatively inexpensive have been used in precursor missions to the outer solar system. Pioneers 10 and 11 (Fig. 12) were intended to explore the environments of

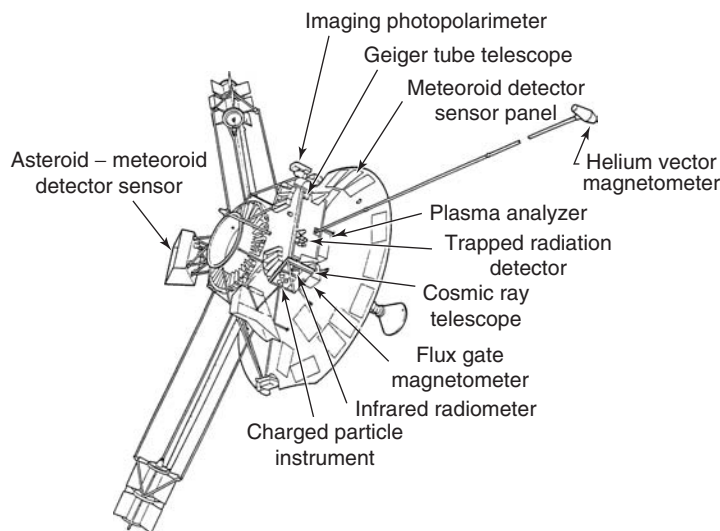


Figure 12. Pioneer 10 and 11 spacecraft: Pioneer 10 was launched in 1972 and flew past Jupiter in 1973. The spacecraft measured the radiative field, the magnetic field, and other elements of the environment to prepare the way for the Voyager spacecraft that was launched in 1977 (see Fig. 2). Pioneer 10 was the first spacecraft to fly past the asteroid belt, to fly past Jupiter, and the first-human-made object to leave the solar system as defined in extent by the orbit of Pluto in 1983. Pioneer 11 was launched in 1973 and was the first spacecraft to fly past both Jupiter and Saturn. In 1979, Pioneer 11 passed between Saturn and the ring plane just outside the A-ring and its visible ring to ensure that Voyager 2 could make a thorough investigation of the Saturn system during its flyby in 1981.

Jupiter and Saturn through which the more sophisticated Voyager spacecraft would have to pass. For spinners, the gyroscopic action of the rotating spacecraft mass is the stabilizing mechanism. Propulsion-system thrusters are fired only occasionally to make desired changes in the spin-stabilized attitude.

Spinners have also been used to launch probes into planetary atmospheres. In this case, the mother ship imparts the proper attitude and spin rates. Examples of this are the Pioneer Venus Multiprobe Spacecraft (Fig. 13) and the Galileo Probe (24) into the atmosphere of Jupiter (Fig. 10). Finally, spinners can be used effectively for radar-altimetry mapping missions such as the Pioneer Venus Orbiter (1978) and the recently completed Lunar Prospector Mission (Fig. 14).

Alternatively, a spacecraft may be designed for active three-axis stabilization. One method is to use small propulsion-system thrusters to nudge the spacecraft back and forth continuously within a dead band of allowed attitude error. Voyagers 1 and 2 have been doing this since 1977 and have consumed about 70 kg of their 106 kg of propellant as of April 2001. They are using propellant at the rate of about 7 grams per week (6). Another method is to use electrically powered reaction wheels, as on Mars Global Surveyor and the recently launched Mars Odyssey. Reaction wheels provide a means to trade angular momentum back and forth between spacecraft and wheels. To rotate the vehicle in one direction, you spin up the proper wheel in the opposite direction. To rotate the vehicle back, you slow the wheel. Excess momentum that builds up in the system due to external torques, caused for example by solar-photon pressure or gravity gradient, must be occasionally removed from the system by applying torque to the spacecraft and allowing the wheels to acquire a desired speed under computer control. This is done during maneuvers called momentum desaturation or momentum-unload maneuvers.

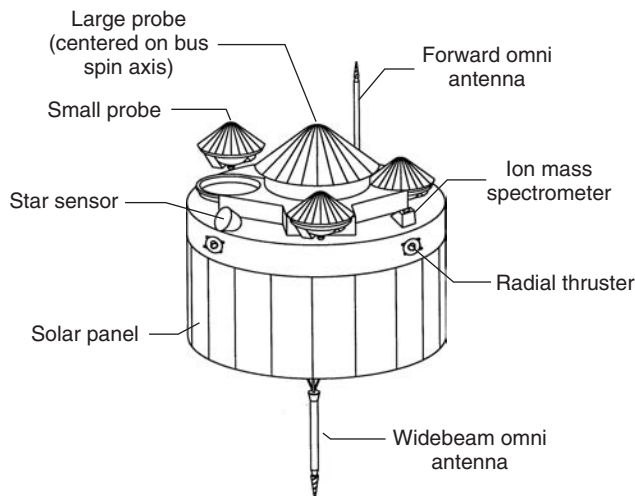


Figure 13. The Pioneer Venus Multiprobe Spacecraft was launched in 1978. It investigated the Venusian clouds and atmosphere by releasing one large probe and three small probes to sample the structure and composition of the clouds, winds, chemical composition, temperature, density, pressure, infrared radiation, and the planet's interaction with the solar wind.

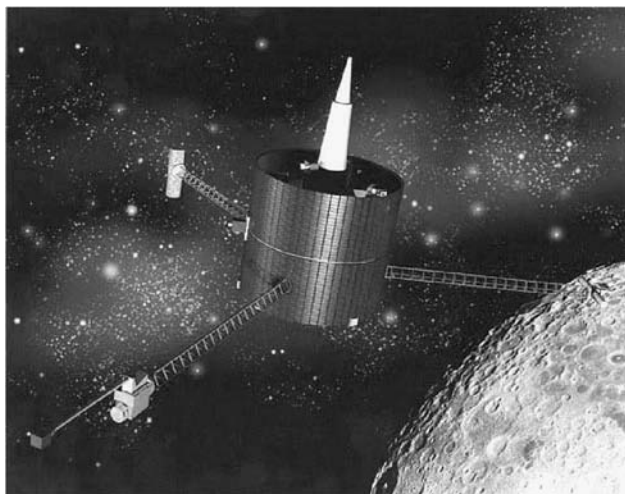


Figure 14. Lunar Prospector's controlled crash into a crater near the South Pole of the Moon in 1999 produced no observable signature of water, but from Prospector's data collected during the mission while in orbit, scientists estimate that up to six billion metric tons of water ice may be buried in craters near the Moon's South and North Poles. Mission findings were also used to develop the first precise gravitational map of the entire lunar surface.

Telecommunications. The telecommunications subsystem provides communication to and from Earth for spacecraft navigation, commanding, and telemetry. Its hardware includes antennae, transponders, and amplifiers. The key functions are signal modulating, demodulation, and some types of data encoding. The key performance parameter is the number of bits per second (bps) that can be transmitted. Among other parameters, the data-rate performance is proportional to transmitter power, the area of the transmitting and receiving antennae, and the square of the transmitting frequency. Planetary-exploration spacecraft use the same Ka-, S-, and X-frequencies as other spacecraft, but there are special frequency allocations for deep space that differ from those for Earth-orbiting spacecraft (25). During the past 25 years, there has been a trend toward higher frequencies for spacecraft telecommunications, from S- to X- and now to Ka-band (32 GHz). Researchers are developing optical communication systems that provide greatly increased performance because of their higher frequency. The current state-of-the-art interplanetary data transmission rate over the Mars range (~ 2 AU) is about 85 kbps, but currently NASA is considering a Mars mission in 2005 that may use a data rate as high as 2 Mbps.

In addition to the primary function of transmitting data to Earth and receiving commands, the telecommunications subsystem is an important element in the navigation system for PES and sometimes is used as part of a scientific investigation. Because of the extreme distances across which planetary spacecraft must be navigated, highly accurate radiometric data are required. To obtain typical accuracy requirements of a few meters for range measurements and 0.1 mm/s for Doppler, a well-calibrated, coherent-ranging transponder is needed. Radio-science experiments use the spacecraft radio and the DSN together as

their instrument, rather than using only an instrument aboard the spacecraft. Radio-science experiments record the attenuation, scintillation, refraction, rotation, Doppler shifts, and other direct modifications of the radio signal as it is affected by the atmosphere of planets, moons, or by structures such as planetary rings or gravitational fields. From these data, scientists can derive a great deal of information such as the structure and composition of an atmosphere and particle sizes in rings.

The receiver-acquisition-and-tracking characteristics are unique for planetary spacecraft. Compared to those of typical Earth orbiters, the receiver threshold is very low, the loop bandwidth is very narrow, and the received-signal strength at the ground station is very low, typically as low as -150 dBm. The DSN provides the largest, most sensitive ground stations available in the world, but closing the telecommunications link (both up and down) is still so difficult that the margin left in the link for planetary spacecraft is far less than the typical 10 dB used on Earth orbiters. Planetary-exploration spacecraft use high-gain antennae for communications; they must be pointed to within a small fraction of 1° . The stringent requirements on the telecommunications link for planetary spacecraft have caused the development of sophisticated channel-coding techniques. Coding is a technique using logic and mathematics to help ensure error-free data transmission. One coding scheme that most interplanetary spacecraft use is a forward-error correction scheme called convolutional coding with Viterbi decoding. Another coding technique is Reed–Solomon coding, which adds bits, and is generally imposed before the convolutional code. A recent advance in coding is called turbo code, which may be used on a mission to Mars in 2005.

Power. The power subsystem generates, conditions, controls, and distributes on-board power. Power sources may include solar-cell arrays, RTGs, or batteries. Electrical power must be conditioned for particular end users, distributed, made stable, and safely controlled to protect the entire system from power failure. Experience shows that a PES requires a few hundred watts up to about 1 kW of electricity to power all of the computers, radio transmitters and receivers, motors, valves, heaters, data storage devices, instruments, a host of sensors, and other devices.

For a PES going beyond about the orbit of Jupiter, the first power-design question is whether to use RTGs or solar arrays. This decision is based upon the Sun range of the mission, the state of readiness of the various power-source options, the acceptable level of mission risk, safety considerations, and the amount of resources (time and dollars) available to the mission. If RTGs are chosen, spacecraft design dramatically departs from that of Earth-orbiter design. Currently, RTGs contain several kilograms of an isotopic mixture of radioactive plutonium in the form of an oxide pressed into a ceramic pellet. The primary constituent of these fuel pellets is the plutonium 238 isotope. Plutonium 238 emits low-energy alpha particles (which can be stopped by a sheet of paper) and very small amounts of gamma radiation. This, coupled with its relatively long half-life (87.7 years) and high melting point (~ 2300 K), makes it the isotope of choice, even though it is expensive and has a much lower power density (4 W/g) than isotopes of cesium, cobalt, and polonium.

The natural radioactive decay of plutonium produces heat (RTGs do not use fission or fusion), some of which is converted into electricity by an array of thermoelectric thermocouples. Thermoelectric conversion does not employ

moving components but instead uses the phenomenon of temperature-induced current flow in materials (Seebeck effect). In this system, a pair of dissimilar semiconductor materials (silicon and germanium), where each end is at a different temperature, is joined in a closed circuit to produce a voltage. Unfortunately, the conversion efficiency is quite low (typically 5%), and RTGs radiate to space about 20 watts for every watt of electric power produced. The spacecraft must accommodate this thermal radiation as well as the nuclear-radiation environment. Although gamma rays and neutrons from RTGs pose no SEE threat, they contribute a major part of the total dose to nearby electronic components. These radiation forms are considered “unshieldable” in the sense that no amount of shielding that would be practical on a spacecraft is effective against these particles. Often, some of the IR sensing scientific instruments can be “blinded” by the warm RTG, and a RTG shade is required. Sometimes a cooling system must be included to keep the spacecraft from overheating when it is enclosed in the launch vehicle. The RTG probably must be conductively isolated from the rest of the spacecraft unless the designer tries to use the RTG waste heat to keep the spacecraft warm, which is a good idea, but makes the thermal design of the spacecraft much more complicated than that of an ordinary Earth orbiter.

Besides complicating the design, RTG-powered spacecraft must be compliant with important environmental and safety issues. The National Environmental Policy Act, for example, requires that an impact statement address nuclear-safety risks. Earth-gravity-assist missions must mitigate the possibility of inadvertent Earth reentry and prove that those risks are negligible.

Propulsion. The propulsion system provides thrust that is used to maintain three-axis stability, control spin, execute maneuvers, and make minor adjustments in trajectory. The propulsion subsystem includes the engines, propellant tanks, pressurant tanks, and associated valves and plumbing. The larger engines may be used to provide the large torques necessary to maintain stability during a solid-rocket motor burn, or they may be the only engines used for orbit insertion. Smaller thrusters that generate between less than 1 N and 10 N are typically used to provide the ΔV for interplanetary trajectory-correction maneuvers, orbit-trim maneuvers, reaction-wheel desaturation maneuvers, or routine three-axis stabilization or spin control.

The requirements and constraints of planetary-exploration missions make the propulsion subsystem designer pay attention to several special issues. Because navigation is so difficult for planetary missions, unmodeled accelerations, such as those from outgassing and leakage must be avoided, and thrusters should be arranged to provide attitude-control torques in coupled pairs to avoid any translational ΔV . For some missions, scientific objectives drive the choice of propulsion system. Sensitive optical surfaces, cryocooled sensors, and low-temperature thermal radiators are vulnerable to contamination produced by the propulsion system. Spacecraft using such devices require propulsion systems that do not exhaust condensable species that can accumulate on these sensitive surfaces. In some cases, a particular propulsion option is eliminated because a spacecraft instrument is designed to detect the same chemical species that are present in the propulsion-system exhaust.

Trajectories used for some planetary missions often require large maneuvers en route these maneuvers use the primary propulsion system many times

in a series of large “burns” required for orbit insertion; insertion can occur years after launch. The primary engine used to deliver the mission ΔV on a planetary spacecraft may have to operate for up to 10 hours and perform 200 cycles, compared to a similar engine used on an Earth orbiter that may need to operate only for 2 hours and perform 5 cycles. For a planetary spacecraft, the propellant supply and pressurization system may have to last for many years. Earth orbiters, such as GEO-communication spacecraft, use most of their propellant during the first few weeks of the mission. After that, the bipropellant main engine and pressurization system are isolated for the remainder of the mission. The long-life requirement for the propellant supply and pressurization system for planetary spacecraft requires careful consideration of pressurant leakage and interactions between propellant and tankage material, which can lead to blocked propellant lines. To reduce the probability of propellant-line blockage, the Galileo spacecraft bipropellant subsystem is activated routinely to reduce the accumulation of corrosion products (26). The requirement for long storage of the propulsion subsystem in space can drive the requirement for special isolation between the propellants that require using additional pyro-and-latch valves. Lack of proper design and isolation within the bipropellant pressurization system on the Mars Observer spacecraft, was probably a leading cause of its failure (27). The Magellan spacecraft, which orbited Venus, used a large solid-rocket motor (SRM) for orbit insertion. The long duration of “storage” in space caused concern whether the SRM would ignite and burn properly. Earth-orbiter SRMs are fired usually within days of launch.

Because of the large ΔV required for many planetary-exploration missions, electric propulsion using ion engines has long been viewed as an attractive alternative to chemical propulsion. Ion propulsion using xenon as a propellant has been demonstrated on the NASA/JPL Deep Space 1 spacecraft (28) (Fig. 15). Ion propulsion (29,30) provides an exhaust velocity about 10 times larger than that of a traditional chemical-propulsion subsystem, and therefore it uses only about one-tenth the amount of propellant to provide the same ΔV . There is a significant savings in propellant, but electric-propulsion systems require additional power-generating and conditioning equipment that offsets to some degree the savings in propellant mass. Interplanetary spacecraft using ion-propulsion systems can undertake the very high ΔV missions that would be prohibitively massive and therefore expensive using traditional chemical-propulsion techniques.

PES Principles and Margins

The preceding pages have described the key requirements and constraints on planetary-exploration spacecraft and the way those requirements and constraints influence the design of the spacecraft. Although designing planetary-exploration spacecraft is a relatively young art, the practice of this art and the struggle against the requirements and constraints for 40 years and about 20 design cycles have identified some detailed technical design principles (31). These design principles, which are being implemented at NASA’s Jet Propulsion Laboratory, originate in the need for very reliable planetary-exploration spacecraft. This section also includes design margins (31) that have been developed to

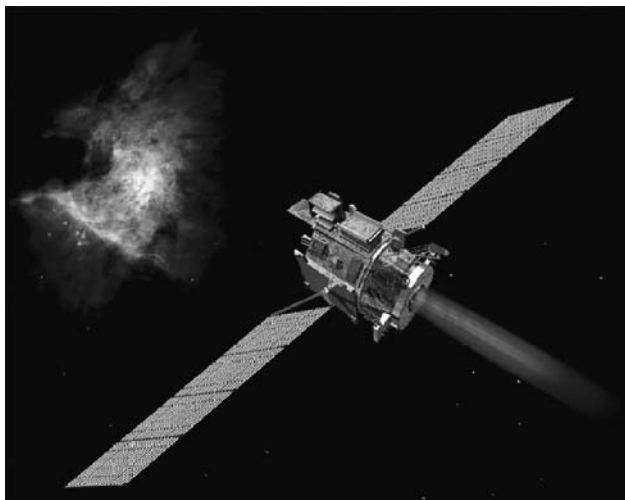


Figure 15. Launched on October 24, 1998, Deep Space 1 was the first mission under NASA's New Millennium Program, designed to test new technologies for future space and Earth-observing missions. During the primary mission, Deep Space 1 tested 12 advanced technologies and instruments aimed at making future science spacecraft smaller, less expensive, more autonomous, and capable of more independent decision-making so that they rely less on tracking and intervention by ground controllers. In a mission extension, the spacecraft flew by Comet Borrelly in September, 2001, returning the best images and other science data ever from a comet. The mission concluded on December 18, 2001. Sources: <http://spacelink.nasa.gov/NASA.Projects/Space.Science/Solar.System/Deep.Space.1/index.html>; <http://nmp.jpl.nasa.gov/ds1/>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

increase the probability that the spacecraft will be ready to launch (with a high degree of certainty) during the inflexible launch window imposed by solar-system celestial mechanics. In many cases, these principles have been developed in response to past failures (27,32,33).

The following principles applied during the design phase help ensure that long-lived spacecraft successfully complete very challenging missions:

- No single electrical and/or mechanical failure shall result in the loss of the entire mission.
- Routine, in-flight power cycling of critical hardware shall be avoided.
- Mission-critical data (flyby science, orbit insertion, etc.) shall simultaneously be recorded onboard and transmitted.
- Thermal design shall keep piece-part, silicon-junction temperatures below 110°C (assuming a 70°C mounting-surface temperature) for circuit design and packaging.
- Electronic hardware shall survive power on/off cycling, temperature cycling, and/or solar exposure cycling of three times the number of worst-case expected mission cycles with worst case flight-temperature excursions. Until a mission estimate is available, the equivalent of 10,000 cycles with a 15°C ΔT (change in temperature) for new/inherited design hardware shall be used.

- The prime-power distribution hot and return lines shall be DC-isolated from the spacecraft chassis by at least 2 ohms. (This ensures a single fault so that a chassis short anywhere in the distribution system among power source, electronics, or user loads does not pose a catastrophic failure.)
- Prime-power on/off switching of electrical loads shall be done by “simultaneously” switching both hot and return sides. This ensures total load removal (no possible ground-return sneak paths) in case of power-related faults.
- Mission-critical deployable design (for example, solar arrays) shall demonstrate a margin of at least 100% under worst-case conditions, particularly cold, stiff cable bundles, vacuum versus air, and coefficient of friction effects.
- Mission-critical separation design (for example, launch vehicle, probe release) shall demonstrate a margin of at least 100% under worst-case conditions.
- Mission-critical mechanisms and actuator design shall demonstrate at least 100% margin for range-of-motion and the end-of-life stage under worst-case conditions, including restart from within any range-of-motion position.
- All electronic-parts radiative capability shall be at least twice that of the expected end of nominal-mission environment.
- A minimum, prelaunch power-on operating time shall be established for all electronics as follows:
 - Unit level prior to spacecraft integration: each electronic assembly, including each side of a block-redundant element, shall have at least 200 hours of operating time.
 - System level before to launch: each single-string electronic assembly shall have 1000 hours of operating time. Each side of a block-redundant element shall have at least 500 hours of operating time and a goal of 1000 hours.

Planetary-Exploration Spacecraft Operating Principles. These principles address the most common and critical issues for operating a planetary-exploration spacecraft far from Earth for many years that has only a single opportunity to perform its unique mission.

- All flight-command sequences shall be tested on a high-fidelity, flight-like system test bed, and all anomalies shall be understood and corrected before sequence-uplink transmission.
- After initiation, mission time-critical operations shall not require “ground-in-the-loop” commanding to enable successful operation/completion.
- Launch-sequence completion shall leave the spacecraft in a ground-commandable, safe state that requires no “immediate” time-critical ground commanding to assure health/safety.
- After in-flight turn on, the downlink-RF transmitter shall not be turned off during nominal flight operations but shall remain powered during the entire mission, except for momentary power cycling via system-autonomous, fault-protection responses.
- Power cycling of mission-critical hardware shall be avoided.
- Prime-selected hardware elements shall remain in use for all operations.

- Swapping to redundant hardware elements shall be limited to fault-recovery actions to ensure health/safety.
- Stored critical data (for example, launch, fly-by science, orbit insertion, entry/descent and landing) shall be protected from loss in the event of selected anomalies (for example, transient power outage) and shall be transmitted to Earth as soon as practical.
- Mission-critical events (for example, launch-vehicle separation and deployments) and verification of deployables shall be available via real-time telemetry.
- Telemetry and command capability shall be available throughout the mission in normal cruise-pointing attitude and during special cruise-phase mission/system activities (for example, long-duration, deep-space trajectory correction, maneuvers, and mission-critical propulsion pyrodevice actuations).
- During non-mission-critical cruise periods following a fault condition, the flight-protection response shall at least autonomously configure the spacecraft to a safe, quiescent, ground-commandable state, preferably transmitting engineering status, but at least an RF-carrier downlink signal.
- During critical-mission activities (for example, launch and orbit insertion), the flight fault-protection response shall autonomously reestablish the needed spacecraft functionality to permit safe, reliable, and timely completion of mission-critical activity.

Design Margins. Design margins are intended to keep the spacecraft design, build, and test program on track to meet the launch date. Design margins must be large enough to accommodate design uncertainties/unknowns and still enable design changes with minimal systemwide effects.

- Spacecraft system-level mass margin shall be at least 30% at the project start; 20% at project preliminary design review (PDR), 10% at critical design review (CDR); 5% at assembly, test, and launch operations (ATLO) readiness; and 2% at launch.
- The spacecraft-system-level power margin for cruise, mission-critical, and safing modes shall be at least 30% at project start, 20% at project PDR, 15% at CDR, and 10% at ATLO start.
- At launch, there shall be at least 10% predicted power margin for mission-critical, cruise, and safing-operating modes.
- At the start of the spacecraft design, the computer throughput and memory capability shall exceed estimated requirements by at least a factor of 4.
- The nominal deep-space link margin shall be at least 3 dB.
- Deep-space links with extreme geometry conditions, surface-to-orbit links, or surface-to-surface links shall consider margins of 10 dB or more, depending on the nature, complexity and scope of design uncertainties.

ACKNOWLEDGMENT

The Jet Propulsion Laboratory (JPL), California Institute of Technology, carried out the research described in this article, under a contract with the National

Aeronautics and Space Administration. The review and contributions of many individuals at JPL are acknowledged, especially Neil Yarnell, John Slonski, David Doody, and Matthew Landano. The efforts of Tom Wilson as editorial assistant are also gratefully acknowledged.

BIBLIOGRAPHY

1. Doody, D. JPL D-20120, Basics of Space Flight, Jet Propulsion Laboratory, Pasadena, CA, 1993. [online], <http://www.jpl.nasa.gov/basics/index.html>.
2. NASA, Mariner-Venus 1967, Final Project Report, (1971) NASA SP-190, Washington, D.C.
3. NASA, Chronology of Lunar and Planetary Exploration, NASA Goddard Space Flight Center, Greenbelt, MD [online], <http://nssdc.gsfc.nasa.gov/planetary/chronology.html>.
4. Kohlhasse, C. (ed.). The Voyager Neptune Travel Guide. JPL Publication 89-24, Jet Propulsion Laboratory, Pasadena, CA, 1989.
5. Jones, C.P., and T.H. Risa. AIAA "The Voyager Spacecraft System Design"; Paper 81-0911, Reston, VA, 1981.
6. Voyager Home Page [online], <http://www.jpl.nasa.gov/missions/current/voyager.html>.
7. Viking Missions Fact Sheet [online], <http://www.solarviews.com/eng/vikingfs.htm>.
8. Cassini Home Page [online], <http://www.jpl.nasa.gov/cassini/>.
9. MESSENGER, a NASA/APL spacecraft [online], <http://sse.jpl.nasa.gov/whatsnew/news-msgr.html>.
10. Near Earth Asteroid Rendezvous (NEAR) Spacecraft Home Page, [online], <http://near.jhuapl.edu/>.
11. Jordan, J.F., and L.J. Wood. Navigation, space mission. *Encyclopedia of Physical Science and Technology*, Vol. 8. Academic, New York, 1987, pp. 742-767.
12. Deep Space Network/Flight Projects Interface Design Handbook. JPL 810-005. Jet Propulsion Laboratory, Pasadena, CA.
13. Bate, R.R., D.D. Mueller, and J.E. White. *Fundamentals of Astrodynamics*. Dover, New York, 1971.
14. Lyons, D. Aerobraking at Venus and Mars: A comparison of the Magellan and Mars Global Surveyor aerobraking phases. AIAA/AAS *Astrodynamics Specialist Conf.* Girdwood, AK, Aug. 16-19, 1999.
15. Lyons, D.T., J.G. Beerer, P.B. Esposito, M.D. Johnston, and W.H. Willcockson. Mars Global Surveyor: Aerobraking mission overview. *J. Spacecraft Rockets* 36 (3): 307-313 (1999).
16. Mars Global Surveyor Home Page [online], <http://mars.jpl.nasa.gov/mgs/>.
17. 2001 Mars Odyssey Home Page [online], <http://mars.jpl.nasa.gov/odyssey/index.html>.
18. JPL. Mercury dual orbiter, mission and flight system definition. D-7443, Jet Propulsion Laboratory, Pasadena, CA, 1990.
19. Mars Climate Orbiter Project Page [online], <http://www.jpl.nasa.gov/missions/past/marsclimateorbiter.html>.
20. Rapp, D., C. Kohlhasse, B. Muirhead, K. Atkins, P. Garrison, W. Breckenbridge, R. Stanton, and L. Wood. *Encyclopedia of Applied Physics*, Vol. 2. VCH, Weinheim, Germany, 1991.
21. Stardust Home Page [online], <http://stardust.jpl.nasa.gov/>.
22. Potts, D.L. The Mars Observer Spacecraft. AIAA, 89-0255, Reston, VA, 1989.

23. JPL. Cassini Project Mission Plan. Jet Propulsion Laboratory, Pasadena, CA, March 12, 1993, PD 699-100-2 Rev B.
24. Landano, M.R., and C.P. Jones. The Galileo Spacecraft System Design. AIAA 83-0097, Reston, VA, 1983.
25. Radio Regulations. International Telecommunications Union, Geneva, Switzerland, 1998.
26. Garrison, P., and C. Jennings. Propulsion systems for long life autonomous spacecraft. AIAA/SAE/ASME/ASEE 28th Joint Propulsion Conf, Nashville, TN, July 6–8, 1992.
27. JPL. Mars Observer loss of signal: Special Review Board Final Report, Publication 93-28. Jet Propulsion Laboratory, Pasadena, CA, Nov. 1993.
28. Deep Space 1 Home Page [online], <http://nmp.jpl.nasa.gov/ds1/>.
29. Jahn, R.G. *Physics of Electric Propulsion*. McGraw-Hill, New York, 1968.
30. Stublinger, E. *Ion Propulsion for Space Flight*. McGraw-Hill, New York, 1964.
31. JPL. Design, Verification/Validation and Operations Principles for Flight Systems. D-17868, Rev. A. Jet Propulsion Laboratory, Pasadena, CA, 11/15/00.
32. JPL. Report on the Loss of Mars Polar Lander, Deep Space 2 Missions. D-18704. Jet Propulsion Laboratory, Pasadena, CA, 3/22/00.
33. Report on the loss of the Mars Climate Orbiter Mission, JPL D-18441. Jet Propulsion Laboratory, Pasadena, CA, 11/11/99.

ROSS M. JONES
California Institute of Technology
Jet Propulsion Laboratory
Pasadena, California

PLASMA THRUSTERS

Plasma thrusters are a class of electric propulsion in which the working medium has the form of plasma in the acceleration zone. The presence of plasma, both positive ions and negatively charged electrons, in the acceleration gap distinguishes plasma thrusters from ion thrusters where the acceleration gaps contain only positively charged ions. In ion propulsion, the space charge field restricts the emissions of ions from the emitter (ion generator), and thus ion thrusters have a relatively low thrust density and require a high acceleration voltage. As a result, they can be efficient only with an acceleration voltage > 1 kV and exhaust velocities of ≥ 30 km/s.

Because the acceleration gaps in a plasma thruster contain both positive ions and electrons, no space charge is needed. For this reason, there are no limitations in theory on the thrust density in plasma thrusters, and the exhaust velocity may range from a few km/s to hundreds or more km/s. Of course, different plasma thruster designs are optimal for each range of exhaust velocities and power levels.

Plasma thrusters, like other electric propulsion thrusters, are of interest for space technology primarily because of their capacity to reach high exhaust velocities, greater than those attained by thermochemical (liquid or solid) fuel rocket engines. In a number of instances, other features of such thrusters may also be of interest.

The chemical energy contained in the working medium is not sufficient to attain high velocities. For this reason, plasma thrusters, like other rocket engines, require special sources of energy. Today, these sources may be either the Sun or nuclear processes of one kind or another (fission, radioactive decay). For each specific flight, optimal velocity parameters exist, defined with respect to minimum total mass of propellant and power source. The optimal exhaust velocities for the majority of purposes are currently within the range of approximately 15–30 km/s.

On the Classification of Plasma Thrusters

These thrusters are classified primarily on the basis of the mechanism for accelerating the plasma. Here, we identify three basic types of plasma thrusters.

- Thermal plasma thrusters in which the electroconductive plasma is heated to high temperatures ($> 3000^{\circ}\text{C}$) by a current passing through it and then is exhausted from the thrusters. Depending on conditions, the positive ions and electrons may be heated at the same time (isothermic plasma thrusters), or predominantly electrons will be heated (nonisothermic plasma thrusters).
- Electromagnetic (Ampere's force) plasma thrusters, in which the plasma is accelerated by Ampere's force:

$$F = BJl, \quad (1)$$

where F is measured in newtons, B is magnetic induction in teslas, J is current in amperes, and l is the length of the current lines in the accelerated plasma.

Because of the limits on power sources in space, the plasma thrusters of today, are developed to generate ones, tens, and hundreds of grams of thrust, although, as noted before, there are theoretically no limits on the amount of thrust that can be created by a single module.

Clearly, the acceleration of conducting plasma by Ampere's force is analogous in nature to the operation of various electric motors. However, in plasma thrusters, the plasma is not accelerated through rotation, but driven in a line, as in linear electric thrusters. Thermal effects in such plasma thrusters have only a slight effect on thrust. It is noteworthy that in the midnineteenth century, Maxwell demonstrated that the effect of Ampere's force may be treated as the pressure of a magnetic field.¹

- Finally, the third class encompasses plasma thrusters in which acceleration is achieved through both kinetic gas pressure and Ampere's force.

The division of plasma thrusters into three groups on the basis of the acceleration mechanism is very general. However, detailed classification of all possible plasma thrusters on the basis of their physical or design features is virtually impossible,

¹It is easy to feel this pressure when you bring the identical poles of two magnets close together.

and we will not attempt to do so here. Instead, we will describe only representatives of the three types of plasma thrusters that have already functioned in space. We will not touch on electrically heated thrusters with low exhaust velocities.

Aside from the three types previously noted, we will describe two more promising candidates for future plasma thrusters and their “progenitor.” Ampere’s force generally plays the decisive role in these thruster types.

Pulsed Plasma Thrusters (PPT). PPTs were the first plasma thrusters used for spaceflight (in 1964 on the Zond-2 spacecraft). There are two reasons that the PPT was the first type of plasma thruster to operate in space. First, because prototypes existed, specifically impulse accelerators (“guns”), which first made it possible to produce plasmoids at velocities of many tens of km/s. The second reason was the simplicity of the PPT propulsion system.

Principles Underlying the PPT. The simplest (from the theoretical standpoint) PPT design is the “railvestron” that has an external magnetic field. This device consists of two rails placed inside an external magnetic field and connected to a power source (Fig. 1a). If the rails are linked by a movable conductive cross-connecting jumper, then current enters the circuit, and under the influence of Ampere’s force, the jumper begins to accelerate.

The accelerator described may be simplified if discharge currents are large (and thus pulses are short). In this case, the device does not need an external magnetic field and runs on the intrinsic magnetic field of the circuit. Such systems have been used to generate plasmoid velocities approaching 100 km/s. However, one complex element still remains in this design, the valve that feeds gas between the rails, whose breakdown creates the necessary plasma connection.

The Bostich (1953) plasma was generated by ablation (destruction and evaporation) of an insulator that separates the electrodes. For this reason, the use of gas valves was rejected fairly early, and PPT virtually always used ablation of one or another insulator, for example, Teflon. Thus, we will deal in what follows only with ablative pulsed plasma thrusters (APPT). Each APPT has two electric circuits: the circuit for igniting the discharge and the main discharge circuit. Ignition of the main discharge uses a small amount of plasma, which is formed, for example, by passing a pulsed current through a contact between the corundum rail and a copper plate. This initiating stage may be located at various points, for example, inside the central electrode, as depicted in Fig. 1b. Plasma enters the main discharge gap through slits in the electrode and induces a discharge close to the insulator, thus causing it to erode. The resulting products of its decomposition are ionized and begin to accelerate because of Ampere’s forces, as well as kinetic gas pressure.

APPT Designs. APPT designs differ depending on the thruster’s purpose. The coaxial design portrayed in Fig. 1b is intended for use with relatively high power (≥ 100 joule). For low power (< 100 J), the preference is for a flat rail design, which was first used on the United States LES-6 satellite (Fig. 1c). Finally, if it is desired to obtain the largest possible impulse for a given power, it is necessary to decrease the exhaust velocity and increase the mass flow rate, that is, switch from an electrodynamic to an electrothermal mode at exhaust velocities of approximately 3–4 km/s. Such a “closed” APPT model was used on Zond-2 (Fig. 1d). Note that two-stage APPT engines have also been developed.

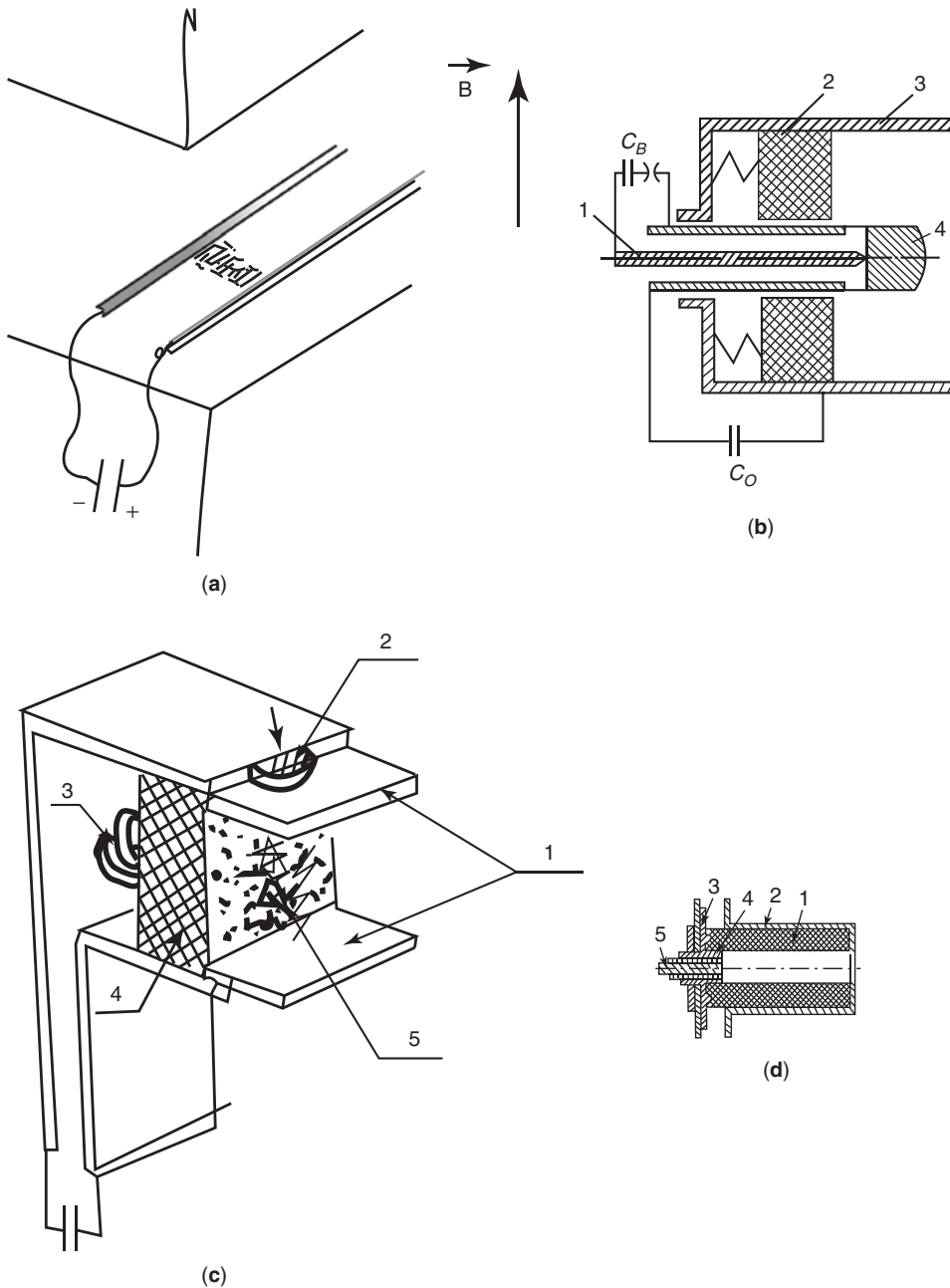


Figure 1. Schematic diagrams of different types of APPTs: (a) railvestron with external magnetic field; (b) coaxial electrodynamic APPT: (1) ignition needle, (2) dielectric, (3, 4) main electrodes; (c) rail type APPT: (1) electrodes; (2) main discharge initiator; (3) spring releasing the dielectric; (4) dielectric; (5) streamer inducing the main discharge; (d) electrothermal coaxial APPT: (1) dielectrics; (2) external electrode; (3) internal electrode; (4) insulator; (5) ignition electrode (Zond-2 spacecraft).

APPT Mode of Operation. Unlike in gas PPTs, in which a previously created cross-connecting jumper is used to provide acceleration, in APPTs during the discharge process and partially after its termination, substances are emitted constantly from the insulator. For this reason, simulation of flow in the APPT requires a hydrodynamic description that takes into account the diverse processes of the ablation of insulators.

Without going into a description of the entire process of discharge in an APPT, we will note the properties of insulator erosion. After breakdown induced by the initiative plasma, the material of the insulator begins to participate in the discharge. This ablation primarily occurs through radiation of plasma, and its intensity correlates well with the momentary power of the discharge. Using small expenditures of energy ($W_c \leq 100$ J), the erosion mass increases as a linear function of W_c , and at high energy expenditure, the erosion mass is proportional to $W_c^{2/3}$. Here W_c is the energy stored in the condensers.

The performance of an APPT increases as the energy expenditure increases in a single pulse. This pertains both to the exhaust velocity, which attains many tens of km/s, as well as to the efficiency coefficient, which reaches several tens of percent at energy expenditures > 100 Joules. Fig. 2 shows how a single impulse depends on energy expenditure in various models.

The factors that restrict the efficiency of an APPT are change in the discharge current over time and vaporization of the insulator at small discharge currents, in which the plasma does not attain the necessary velocity. Discharge nonstationarity depends on correct selection of the insulator material. Insulators made of Teflon have proved to be quite good.

The First APPT Spaceflight. The first flight took place in Russia. Work on pulsed plasma thrust at the Atomic Energy Institute (AEI)² began in late 1960, and by 1961, efforts were already concentrated on ablative pulsed plasma modules. Development work was conducted under the direction of A.M. Andrianov. They developed thrusters with various parameters, including sustainer thrusters for interplanetary flight. In late 1961, contacts were established with the S.P. Korolev Design Bureau, which proposed to develop an electrothermal APPT for the planned Zond-2 spacecraft. Figure 3 shows an electrothermal propulsion system with an APPT (developed with the collaboration of V.A. Khrabrov), which was used for the attitude control system on the Zond-2 interplanetary probe. After initialization of discharge, gas heated to a temperature of approximately 5000°C was exhausted at a velocity of approximately 2–4 km/s, imparting the corresponding impulse to the thruster.

The parameters of this propulsion system with 6 APPTs are as follows: energy stored in the condensers – 56 Joules; operating frequency – 1 Hz; single pulsed thrust – 2×10^{-3} N-s, total thrust impulse – 7.2×10^3 N-s thrust efficiency – $\sim 6\%$; working medium supply – 3 kg; mass of the propulsion system – 28.5 kg; service life – 5×10^5 impulses.

This system was turned on at a distance of ~ 1 million km, and thus did not run very long. However, all at the parameters listed here were confirmed during ground tests.

²Currently the Kurchatov Institute Russian Science Center.

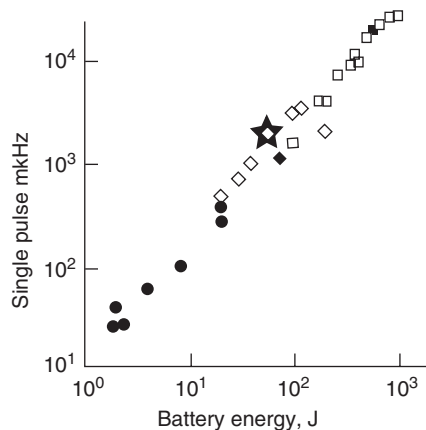


Figure 2. A single pulse (mKHz) as a function of battery power.

Subsequent APPT Development. In the West, the first APPT launch occurred in 1968 for station keeping of the US stationary communications satellite LES-6. It was developed by the MIT Lincoln Laboratory. The energy expenditure on discharge was ~ 2 joules. The specific impulse was 250–500 s. The mode of operation was almost thermal, and the system functioned for approximately 10 years.

On the LES-8/9 satellite launched by the United States in 1973, the energy expenditure on discharge reached 80 joules. The maximum specific impulse was

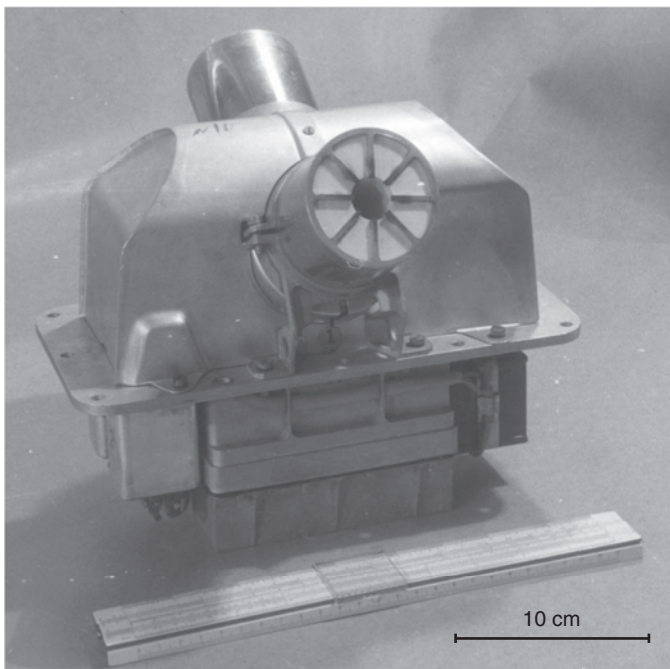


Figure 3. APPT for the Zond-2 spacecraft.

1450 s. Thirty mg of insulator was vaporized on one impulse. This APPT operated in a purely electrodynamic mode.

There are two more APPT spaceflights that are noteworthy. In 1981 China launched MDT-2A. Here, the energy expenditure on an impulse was $W_c = 23.9$ joules. The maximum specific impulse lasted 990 s. Finally, in the fall of 2000, the United States launched the EO-1 satellite containing an APPT. The energy expenditure on an impulse was ~ 60 joules, the specific impulse was ~ 1000 s. The efficiency was $\sim 20\%$ at an energy expenditure of up to 100 joules per impulse. The total mass of the system was 31 kg, of which 12 kg was the insulator, that is, the working medium. This propulsion system was still operating in the fall of 2001.

Future Prospects for APPT. Various APPT models developed in laboratories had a service life of 10^7 impulses and a total impulse of 10^5 N-s. The energy in one discharge ranges from fewer than 10 joules to thousands of joules. The specific impulse increases with discharge energy and reaches 2500 s at an energy expenditure of 200–300 joules.

Interest in such thrusters waxes and wanes, and it is highly likely that with time they will occupy a reliable ecological niche. This optimistic opinion can be justified by three unique (at least at the present) properties of pulsed propulsion systems:

- They do not require preparation time to start operating.
- There are no valves in the erosion version.
- They can generate small impulses ($\sim 10^{-5}$ N-s), which is important for precise attitude control of a spacecraft.

Today an additional argument in favor of APPT is the continually increasing interest in small satellites.

Not even considering future APPTs, we must bear in mind that, from the standpoint of minimal impulse ($\sim 10^{-3}$ N-s) and high efficiency ($\sim > 40\%$), a stationary plasma thruster (SPT) operating in a quasi-stationary mode may be useful. But SPTs require a preparation period before they can start operating (~ 1 minute) and have valves.

Stationary Plasma Thrusters (SPT). After the APPTs, the next PTs that were launched into space (in 1971) were stationary plasma thrusters. These plasma thrusters operated or currently operate on more than 60 satellites. We will concentrate on them for this reason. First, we will provide a brief theoretical introduction.

On the Dynamics of Particles in Electromagnetic Fields. SPTs are plasma thrusters, and thus their operation can be described using formula 1 for Ampere's force and the so-called generalized Ohm's law, considering also the Hall effect. However, the operation of SPTs is more clearly described using the language of electron and ion dynamics.

Several general factors are noteworthy. In the absence of an electric field and at a constant value of \vec{B} in space and in time, the general motion of particles takes the form of a helix, which can be considered a superposition of two motions: along \vec{B} with a constant velocity component V_{II} and revolution around a

circumference in a plane perpendicular to \vec{B} at an angular frequency of ω_H and radius ρ_H :

$$\begin{aligned}\omega_H &= \frac{eB}{M}, \\ \rho &= \frac{V_1}{\omega_H}.\end{aligned}\tag{2}$$

This frequency and radius are called, respectively, the Larmor or cyclotron frequency and radius. If, aside from the homogeneous and constant field \vec{B} , there is another homogeneous and constant electric field perpendicular to \vec{B} , then a more complex motion is generated, which can be considered the superposition of the helical motion described and motion in a direction perpendicular to \vec{E} and \vec{B} at constant velocity,

$$U_E = \frac{E}{B},\tag{3}$$

which is called “electron drift velocity.” The perpendicular electric current that is thus generated is often called a Hall current. The current along the E field (the usual Ohm current) is virtually absent under these conditions. However, this happens only in the absence of collisions with heavy particles. When these particles collide, the electron loses its drift velocity and to gain speed again under the influence of the E field, it is displaced toward the anode. This generates an Ohm current. Its magnitude depends on the frequency of collision, or to be more precise, on the product of the Larmor frequency ω_H and the free path time τ . The physics of SPT — its plasma dynamics for $\omega_e \tau_e \gg 1$. For this reason, SPTs are frequently called Hall accelerators.

Equipotentialization of Magnetic Force Lines. Electrons can freely move along force lines, and this leads to a situation where in plasma, the force lines of a magnetic flux with accuracy up to comparatively weak thermal disturbances become equipotential. For this reason, force lines with superimposed helical electron trajectories can be considered transparent magnetic electrodes, moving at drift velocity. In other words, we can coin the term “electromagnetic electrodes.”

Now let us return to the SPT.

SPT Design. SPTs are depicted in Fig. 4a. Here, we distinguish three modules:

1. a magnetic system consisting of a magnetic conductor (1a), axial symmetrical poles (1b), and a magnetization coil (1c), generating a quasi-radial magnetic field in the channel;
2. a gas-plasma loop, consisting of the gas tube (2a), a gas distributor (2b), a buffer volume (2c), and an axially symmetrical dielectric channel (2d);
3. an electric discharge circuit, including an electron receiver anode (3a), a power source (3b), and a thermocathode electron emitter (3c).

The magnetic system generates a quasi-radial magnetic field, increasing as it approaches the end of the accelerator. The force lines of this field are convex

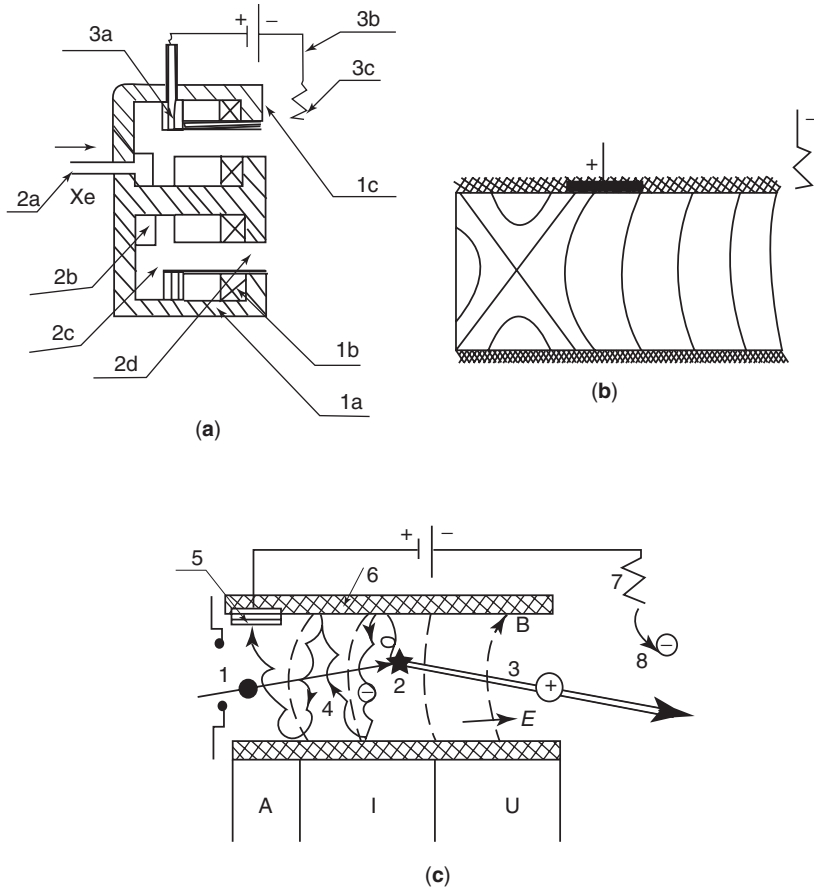


Figure 4. (a) SPT design (explanation in text). (b) The nature of magnetic force lines in the SPT channel. (c) Functional design of an SPT: (1) neutral atom; (2) its ionization; (3) the current that has been generated is accelerated by the electric field; (4) the free electron forming during ionization diffuses on the anode; (5) the anode; (6) the walls of the channel; (7) thermal cathode; (8) electron emitted by cathode; A- anode zone.

toward the anode (Fig. 4b). The magnitude of this field is selected so that the electronic Larmor radius ρ_e is much less than L , the length of the discharge gap, and the ion radius ρ_i is much greater than L because the ions are virtually unaffected by the magnetic field:

$$\rho_e \ll L \ll \rho_i \quad (4)$$

In actuality, today we are discussing magnetic fields at a maximum of $\sim (1.5-3) \times 10^{-2} \text{ T}$, and channel length of $\sim 2-3 \text{ cm}$. When the magnetic field increases as it approaches the end, the plasma configuration is macrostable (for more information, see the next section), and thus a rather stable distribution of the electric field forms in the channel. The voltage of the longitudinal electric field in the main part of the channel is proportional to the magnitude of the radial magnetic field component.

The Functional Design of the SPT. Because of the presence of mutually perpendicular E and H fields, the electrons drift along the azimuth, creating an azimuthal current. Its interaction with the radial magnetic field generates Ampere's force. The presence of azimuthal drift in this model involves "the rotation of the whole system of electromagnetic electrodes (Fig. 4c)."

Neutral atoms emitted from the gas distributor are caught up in the rotating cloud of electrons and are ionized there not far from the anode (Fig. 4c). Because the density of the neutral atoms in this zone is $\sim 3 \times 10^{13} \text{ cm}^{-3}$ and the ion density is $\sim 10^{12} \text{ cm}^{-3}$, collisions among atoms and ions are negligible, but this is not true of their collisions with electrons. For this reason, if an ion generated at a point of potential Φ^* does not collide with the walls and is not ionized a second time, it accelerates under the influence of the E field and leaves the channel with energy $\varepsilon = e\Phi^*$.

Experiments have shown that the mean energy of the ion in the exhaust is

$$\langle \varepsilon_i \rangle \approx (2/3 - 3/4)eU_p, \quad (5)$$

where U_p is the difference in potential between the anode and the cathode. The degree of ionization of the flux of xenon atoms, which today is the main working fluid in the SPT channel is very high, $\sim 95\%$. The effective thickness of the ionization zone is $\sim 4\text{--}5 \text{ mm}$. Although the electrons that form during ionization appear in the magnetic field, they must hit the anode. There are several factors that facilitate this. Having reached the anode one way or another, the electrons then pass through the external circuit and are emitted from the cathode in the ion flux that is exhausted from the SPT channel, providing "spatial," and also "current" neutralization of the ion flux.

The discharge current is a relatively accurate function of mass output. The volt-ampere characteristics in Fig. 5 frequently point to saturation of the

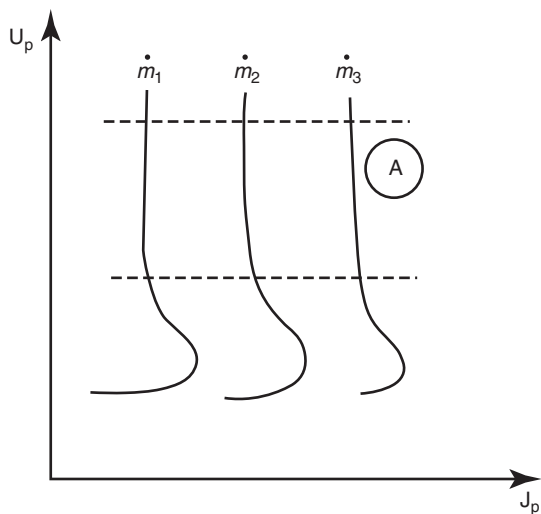


Figure 5. Schematic graph of volt-ampere characteristics for various mass expenditures ($m_3 > m_2 > m_1$).

discharge current for a given output m . Typical values for the thrust efficiency coefficient (taking into account loss in the magnetic system and cathode) are of the order of 50–65%.

Characteristics of the Physical Process in the SPT. From the standpoint of physics, an SPT is a completely new type of gas discharge system, which is unlike all other classic or new discharge devices. We will discuss two theoretically important points: the macrostability of the plasma configuration and the role of electron collisions with the dielectric walls of the channel.

On Discharge Stability in an SPT. For almost 50 years, no one succeeded in generating a superheated electric field in a well-ionized plasma using a magnetic field. As a result, it was concluded that this was impossible (the “Baum barrier”). But the developers of the SPT succeeded in overcoming the Baum barrier by suppressing the large-scale azimuthal nonsymmetrical instability, the so-called “spike,” which had been short circuiting the anode and the cathode within the SPT channel. Later, A.I. Morozov proved that the “spike” disappears if the induction of the magnetic field increases from the anode to the outlet. Specifically the magnetic field that increases as it approaches the outlet is incorporated in the concept of the SPT (see Fig. 4b).

On the Role of Electron Collisions with the Walls of the SPT Channel. SPTs are characterized by high length of free electron paths, the presence of magnetic force lines abutting the channel walls, the presence in the space of a longitudinal electric field, and a high electron temperature ($T_e \geq 15$ eV). The combination of these factors gives rise to a series of specific phenomena:

1. The distribution of electrons in the circuit is, in principle, not Maxwellian.
2. The dispersion of electrons on the walls leads to additional transfer of electrons transverse to the magnetic field—this is called “wall conductivity.”
3. The first experiments showed that the parts of the insulator that adjoined the outlet underwent disintegration because of ion bombardment, which limited the working life of the thruster. Later, during long-term service life tests (> 1000 hours) of the SPT in the Fakel Design Bureau (Kaliningrad), a new form of erosion was discovered. A periodic structure was formed in the insulator, oriented not at right angles to, but along the ion flux (Fig. 8a). The period was of the order of the length of an electron Larmor radius. Evidently, electrons play an important role here; electron impacts on the insulator in one way or another cause it to “disassociate” and isolate individual atoms or small clusters.

The EOL Propulsion System. The idea for an SPT as a system with a strong superheated volume electrostatic field began to be formulated in 1962–1963. By the end of 1963, a study of the first SPT model was started in the laboratory of A.I. Morozov and G.Ya. Shchepkin at the AEI in the department headed by L.A. Artsimovich. By mid-1970 a full-scale mockup of the EOL propulsion system containing an SPT had been built there. Subsequently, it was transferred to the Fakel Design Bureau in Kaliningrad, where the team of Chief Designer, R.K. Snarskiy brought it up to the necessary level; by late 1971, it had been installed in the Meteor satellite.

The propulsion system's major parameters are as follows:

- propulsion system power ~ 400 W
- thrust ~ 0.02 N
- mass ~ 18 kg.

The design of the EOL SPT is depicted in Fig. 6. On 19 December 1971, the Meteor satellite was launched into space.

At the start of 1972, the EOL propulsion system with an SPT passed performance tests. The main results were

1. demonstration of the reliability of the EOL propulsion system with an SPT and of its compatibility with the satellite;
2. no significant effect of plasma structures on communications with Earth;
3. good correspondence between thrust characteristics measured on the test stand and those derived from changes in satellite trajectory. Instead of the announced working life of 100 hours, the EOL system operated for ~ 150 hours, used its entire supply of xenon, and provided useful correction of the satellite orbit, putting it into a conditionally synchronous orbit, and raising the orbit's altitude by 15 km.

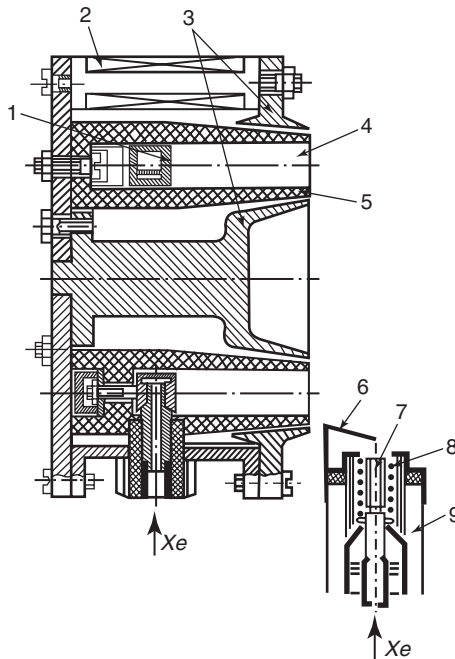


Figure 6. Diagram of the first SPT (SPT EOL): (1) anode; (2) magnetic coils; (3) magnet poles; (4) acceleration channel; (5) insulator; (6) shielded electrode; (7) cathode made of LaB_6 ; (8) start-up heater; (9) cathode-compensator housing.

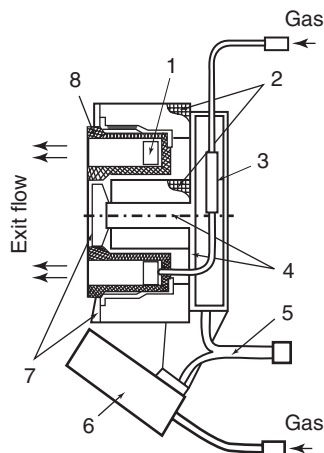


Figure 7. Diagram of the SPT M-100: (1) anode; (2) magnetizing coil; (3) electric isolator in the gas loop; (4) magnetic circuit; (5) electric cable; (6) cathode; (7) magnetic poles; (8) discharge chamber.

Thus, the Fakel Design Bureau had been provided with a useful design for an electric propulsion thruster. As a result, the Design Bureau has been producing propulsion systems with SPTs for more than 30 years. The EOL SPT had an efficiency coefficient of $\sim 35\%$. Soon after the launch of EOL-1, the AEI proved that the efficiency could be increased significantly (to $\sim 50\%$), if there were a shift to “thin” pole strips, that is, if the gradient (H) along the length of the chamber were increased. This was implemented on the EOL-2 propulsion system.

Further Work on SPTs at the Fakel Design Bureau. Further work involved continuing to improve the design of SPTs, as well as correcting the geometry of the magnetic field. The latter fostered improved stability of SPT operation on the test stand but did not increase its efficiency or lead to diminished divergence of the flux ($\alpha_{\text{eff}} \sim \pm 45^\circ$). One of the main ideas that guided the developers of this magnetic system was the desire to displace the acceleration zone maximally to the end of the channel, which lowered efficiency and increased (or in the best case failed to decrease) the angle of divergence.

Nevertheless, in the late 1970s, the Fakel Design Bureau, collaborating with the Moscow Aviation Institute, developed a standard module SPT M-70.³ Its rated power was ~ 1 kW, thrust ~ 4 g and efficiency 45–50%. The M-70 SPT was installed on nearly 40 satellites.

In the early 1990s, development of a new thruster, the SPT M-100, began. Its rated power was ~ 1.5 – 2 kW. It was close in design to M-70, but the M-100 had a thicker insulator near the outlet. A schematic drawing of the thruster is provided in Fig. 7. Its external appearance before starting to operate is shown in Fig. 8a and after 5000 hours in Fig. 8b. The performance characteristics of this model are provided in Fig. 8c. The working voltage for the M-100 is ~ 250 – 400 V, the thrust is 6–10 g, the efficiency is $\sim 50\%$, and the flux divergence is $\alpha_{\text{eff}} \approx \pm 45^\circ$.

³The number 70 indicates the interior diameter of the external channel insulator.

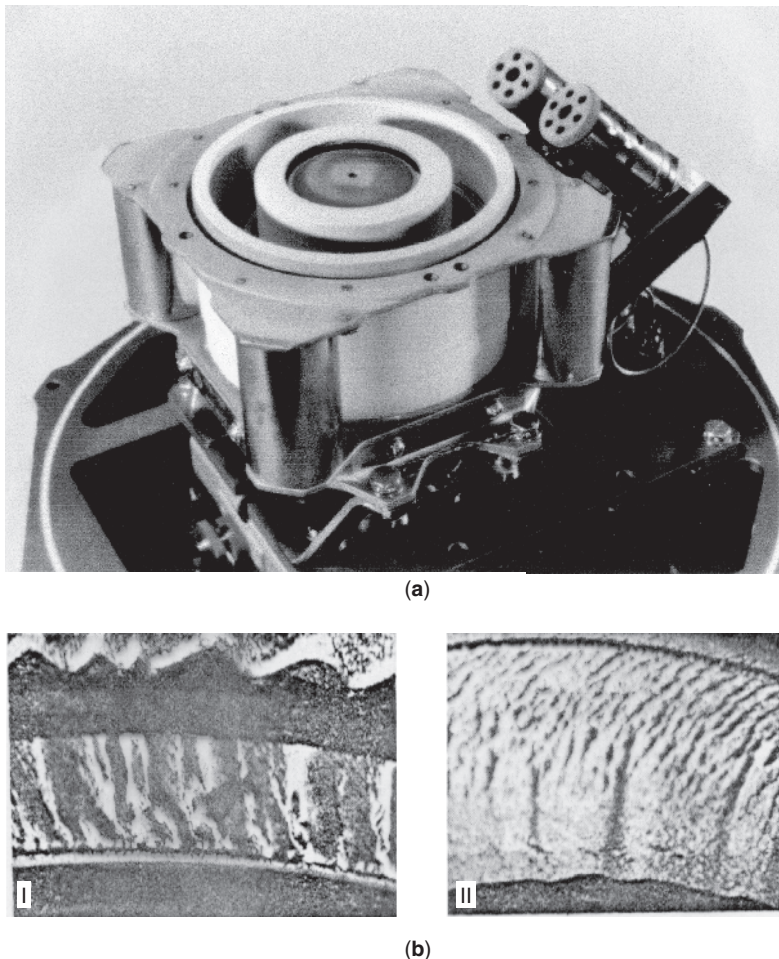


Figure 8. (a) Exterior view of the SPT M-100. (b) Erosion of the SPT M-100 after 5000 hours of service life tests: (I) magnified picture of a fragment of the external insulator; (II) magnified picture of a fragment of the internal insulator. (c) Integral performance characteristics of the SPT M-100 using Xe with mass loss in the cathode of $m_k = 0.4$ mg/s and in the anode of (a) $m_A = 3.6$ mg/s; (b) 4.7 mg/s., and (c) 5.6 mg/s. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The most remarkable property of the M-100 is its working life. Although during the course of 9000 hours, the ion flux “ate” the insulator and also part of the poles, the thrust characteristics remained virtually unchanged. Currently the M-100 is operating on many satellites. The successes of the SPT stimulated similar development by other firms—in Russia, the United States, a number of countries in Western Europe, China, and Japan. It is noteworthy that the SPT and similar systems are attracting increasing attention as new tools for beam processing of material surfaces.

The ATON SPT. The SPTs described before have relatively low efficiency ($\sim \pm 50\%$) and high beam divergence. It became clear that this was associated with nonoptimal conditions in the ionization zone of the working medium and

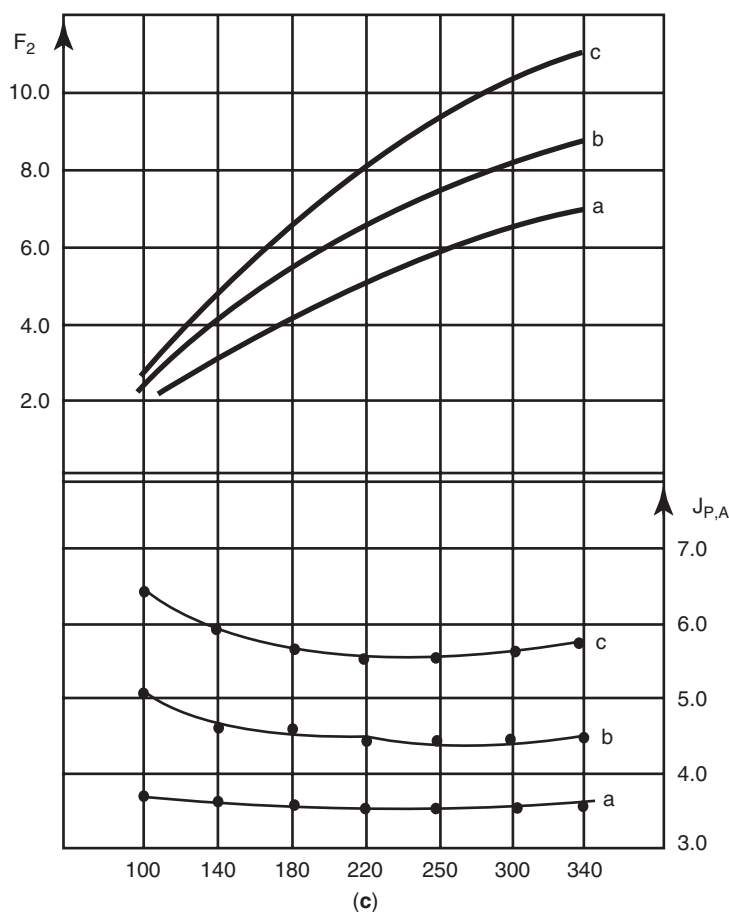


Figure 8. (Continued)

also with nonoptimal geometry of the magnetic field in the channel. It was also clear that divergence of the flux could be radically decreased only if the ionization zone were located not at the outlet of the channel but in the middle in the zone where the curvature of the force lines was the greatest, that is, not far from the zero point of the magnetic field. This is where the anode had to be. As a result, the SPT design shown in Fig. 9 was generated.

The laboratory model of this thruster was called the SPT ATON. The dimensions of the device were approximately the same as those of the M-70. After optimization, this model displayed unique performance characteristics. Its efficiency (taking account of loss of xenon in the cathode compensator) approached $\sim 65\%$, and the effective angle of divergence was $< 10^\circ$. This model was built in the Moscow Institute of Radiotechnology, Electronics, and Automation (A.I. Bugrovaya's laboratory) in collaboration with A.I. Morozov (AEI) and financial support from the SEP company of France.

The Future of SPT Designs. This design permits significant expansion of the range of working parameters. Thus, in the AEI laboratories, models using external insulators with diameters of 300 mm were tested. The power of this

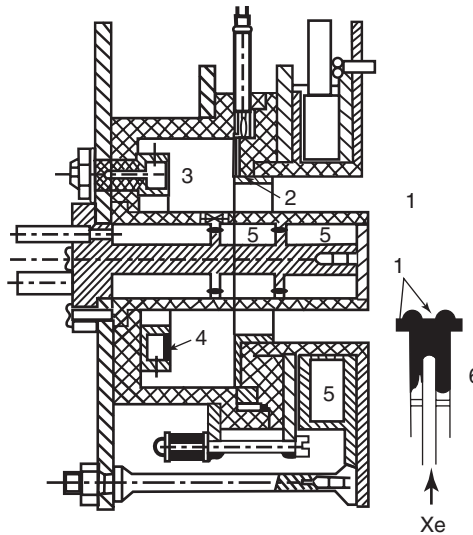


Figure 9. ATON SPT: (1) acceleration channel; (2) anode; (3) buffer volume; (4) gas distributor; (5) magnetic coil; (6) cathode.

thruster reached 30 kW, and the thrust, depending on discharge voltage, fell in the interval of 100–150 g (1–1.5 N). The efficiency under these conditions was $> 65\%$. The working medium was xenon. But this is far from the limit. In a quasi-stationary mode with impulses lasting ~ 10 ms, the so-called “two-lens model” of the SPT produced hydrogen ion fluxes with energy of ~ 5 keV and an ion flux of 5 kA, that is, the accelerator power reached 25 MW, and the streaming velocity was ~ 1000 km/s. Finally, note that SPT tests were performed using various working media, ranging from hydrogen to xenon. Performance characteristics close to the physical limit were obtained.

Anode Layer Thrusters (ALT). There is a whole series of systems called “anode layer thrusters.” However, as of the fall of 2001, only one of these has been launched into space (in October 1999), the so-called hollow-anode ALT. We will discuss this system in more detail here.

Design of the Hollow-Anode ALT. As Fig. 10 shows, this system, in many ways, looks similar to the SPT. But the length of the channel here is minimal, and its walls are not dielectric but metal at the same potential as the anode. The potential drop between the anode and the cathode (the body of the thruster) occurs within the confines of a narrow gap $d \leq 1$ mm. There are three important points to be made here. First, the ionization zone is basically concentrated inside the hollow anode. Second, the drop in potential occurs in the vicinity of the force lines, which are emitted from the gap, and third, this permits moving the acceleration zone outside the bounds of the anode cavity, if the edge of the anode strip is moved to the end of the magnetic poles and the screens covering them. This feature of the design creates the preconditions for a sharp attenuation of the ionic disintegration of the entire exhaust portion of the thruster but at the same time increases the divergence of the ion flux. As for the ionization zone, concentrated in the hollow anode, it seems that here the ionization is caused by electrons of relatively low

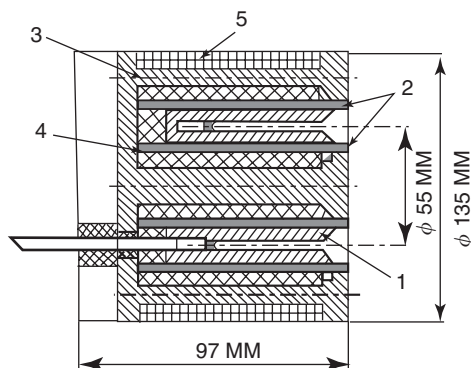


Figure 10. Schematic cross section of ALT with a hollow anode: (1) external magnetic pole; (2) anode; (3) internal screen and internal magnetic pole; (4) magnetic circuit with coils; (5) external screen.

energy. Such electrons may be generated in this hollow space because of the differential in potential, which occurs here as a result of the jump in potential near the walls of the anode. Unfortunately, local measurements of plasma parameters, including potential in the working zone of this thruster, have not been published.

The ALT Going into Space. The design described was implemented in a Russian–U.S. project to build and test such a thruster in space. The flight model of the thruster was developed and manufactured by the Central Scientific and Research Institute of Machine Building, where it was named D-55, the number 55 indicates the exterior diameter of the channel in millimeters. It was tested at this institute and at the Jet Propulsion Laboratory (United States).⁴ The external appearance of the D-55 is shown in Fig. 11.

Figure 12 shows the specific impulse and thrust efficiency as a function of power. It is evident that these characteristics are close to, although somewhat lower than the characteristics of the SPT-100 at various levels of power. Exhaustive life service tests were performed for 600 hours. Erosion of the thruster components was low, and the investigators consider that the full working life will be of the order of ~ 5000 hours. The remaining parts of the propulsion systems were developed and manufactured in the United States and have been flight-tested on the STEX satellite. Preliminary experiments in space were conducted on 23–24 October 1998. The thruster was turned on ten times and operated for a total of 100 minutes. A thrust of approximately 0.04 N was obtained, and orbital altitude was raised by 650 m. Although the results of long-term tests have not been published, it is known that this system implemented its assigned program during the course of a year.

ALT Evolution. The hollow-anode ALT design described was the culmination of almost 40 years of evolution. One precursor, a system for obtaining hot plasma in a trap, was developed by M.S. Joffe (Fig. 13a); ions with energy of several hundred eV were obtained by applying voltage between cold plasma coming from an arc source to the axis of a trap and the exterior body of the trap. Measurements, taken by Ye.Ye. Yusmanov, showed that here the distribution of

⁴In the United States, it was called the TAL-WSG.

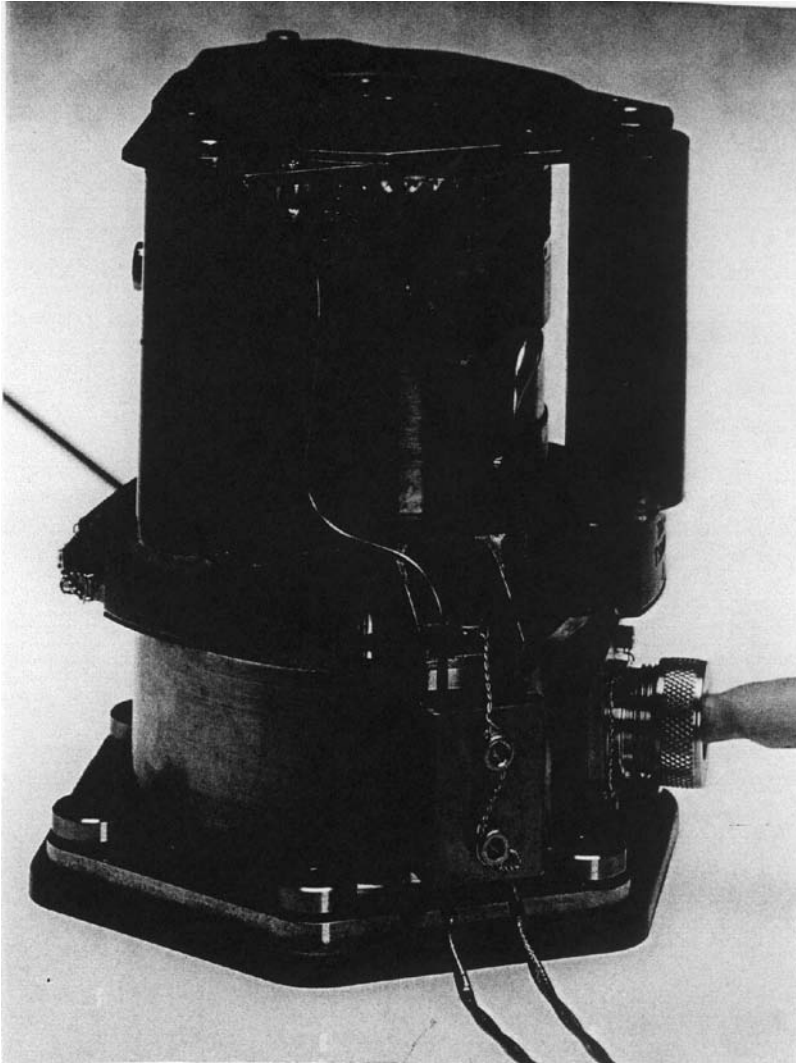


Figure 11. Exterior view of the ALT D-55 (DAL-WSF).

potential was very inhomogeneous. A narrow zone with a high potential differential was observed close to the axis, whereas further away, there was an almost equipotential region.

A theoretical model of this structure was proposed in the late 1950s by A.V. Zharinov, who was then working at the Atomic Energy Institute. He showed that the width of the layer that forms is of the order of the electron Larmor layer and confirmed this by developing a simple “ion magnetron” (Fig. 13b) that produced a radially dispersing flux.

Soon, having transferred to the Central Scientific and Research Institute of Machine Building, he began work to use this principle to develop a plasma thruster. As in the SPT, he used a magnetic field that increased as it approached the end of the channel and obtained a stably burning discharge. This work

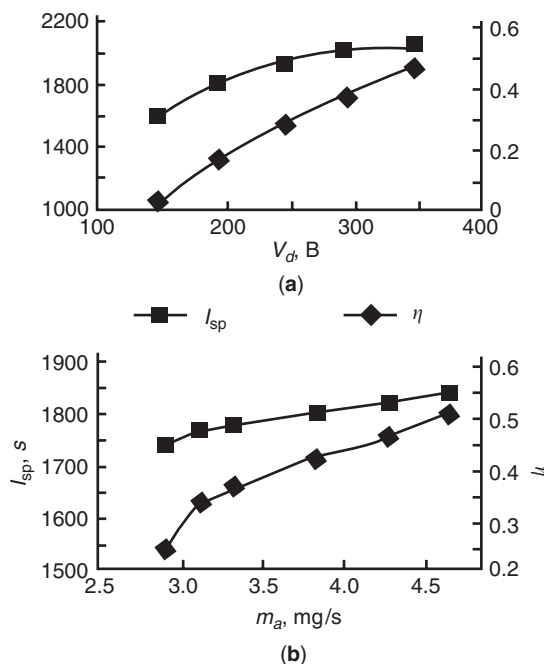


Figure 12. Efficiency coefficient and specific impulse of ALT-55 as a function of discharge voltage and consumption of xenon.

culminated in the 1960s in creation of a two-stage ALT with a “hollow cathode.” One problem it had was that it only operated well at high voltages, $U > 2$ kV, and thus, despite the use of bismuth and lead as the working media, the specific impulse was too high for current uses. The power of this thruster reached 100 kW. Its working life at 16 kW was determined by a 500-hour test.

After a pause in the 1970s, spurred by the successes of SPTs, the Central Scientific and Research Institute of Machine Building began to search for an efficient one-stage model with a metal channel. Finally, in the late 1980s, A.V. Semenko and his colleagues developed an ALT with a hollow anode, which has now flown in space.

The Future of ALT. Today, it is difficult to predict clearly the future of one-stage ALTs with hollow anodes as a universal thruster. For the time being, its performance characteristics are not as good as those of the SPT in efficiency and flow divergence, especially compared to the ATON SPT. The working life is also not clear because extrapolation is very unreliable. Nonetheless, the possibility has not been ruled out that, in the area of high specific impulses ($I_{sp} > 5000$ s) and high power, thrusters in the ALT family may have great advantages. Despite these reservations it is clear that this developmental trend is worth the most serious consideration.

New Plasma Thruster Candidates

Virtually, any design for a plasma accelerator could be used for a plasma thruster, given appropriate optimization. Moreover, a number of systems, starting in the

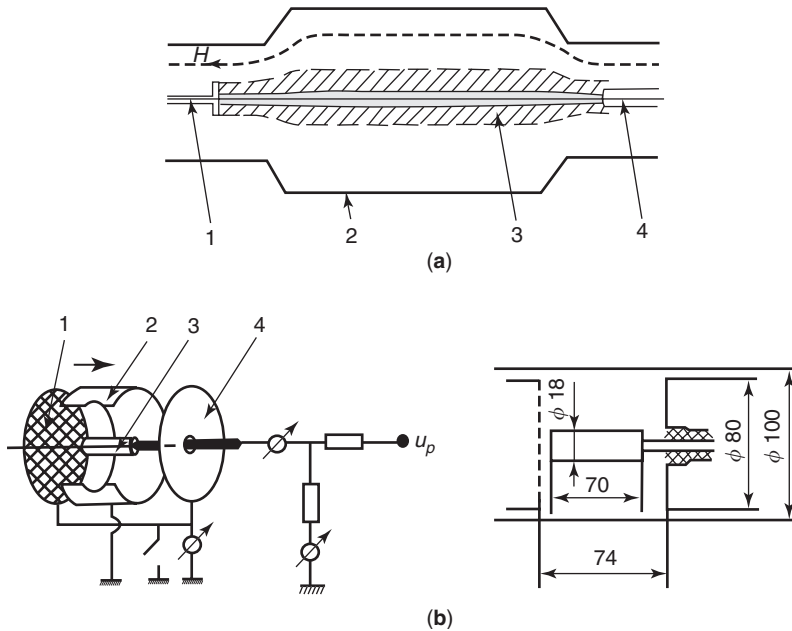


Figure 13. Ion magnetrons: (a) The set-up used by M.S. Joffe and Ye.Ye. Yushmanov: (1) plasma source; (2) chamber walls; (3) plasma beam; (4) reflector ($U_p = 5 \text{ kW}$, $N = 2 \text{ kE}$, $P = 5 \times 10^{-5} \text{ torr}$). (b) A.V. Zharinov's ion magnetron: (1) mesh end cathode; (2) cathode; (3) cylindrical anode; (4) solid end cathode ($U_p = 1\text{--}10 \text{ kEV}$, $N \leq 1.3 \text{ kE}$, $P = 10^{-3}\text{--}10^{-5} \text{ torr}$).

1960s, were optimized as plasma thrusters and even tested on geophysical rockets but were never installed on a satellite for one reason or another.

The greatest amount of attention in the USSR (Russia) has been devoted to the so-called end accelerator—the “high-current end accelerator” and the “Hall end accelerator.” In the West, the latter type of plasma accelerator is called a “magnetoplasma dynamic thruster.” We will describe both of the accelerators briefly, because for many reasons, they are not likely to become multipurpose thrusters in the foreseeable future, like the SPT, although they may prove optimal for special uses.

To facilitate understanding of the ideas underlying these thrusters, it is beneficial to remember the conditions under which they were developed. The euphoria engendered by the first successes in the conquest of space naturally gave rise to new plans. As early as 1959, there was serious discussion of sending a manned spacecraft to Mars. Evaluations showed that this would require a plasma thruster with a thrust of $\sim 10 \text{ kg} = 100 \text{ N}$, specific impulse of $\sim 10,000 \text{ s}$, and, thus, power of $10,000 \text{ kW}$ at an efficiency of $\eta = 0.5$.

Magnetoplasma Analog of Laval's Nozzle. The need for thrusters on this scale naturally suggested the development of electrodynamic systems with intrinsic magnetic fields. The first thruster design of this type was proposed in 1959 in the USSR by A.I. Morozov. This was the idea of a stationary, high-current plasma thruster, representing a magnetoplasma analog of a typical gas dynamic nozzle with a central body (Fig. 14a). The central body would play the role of a

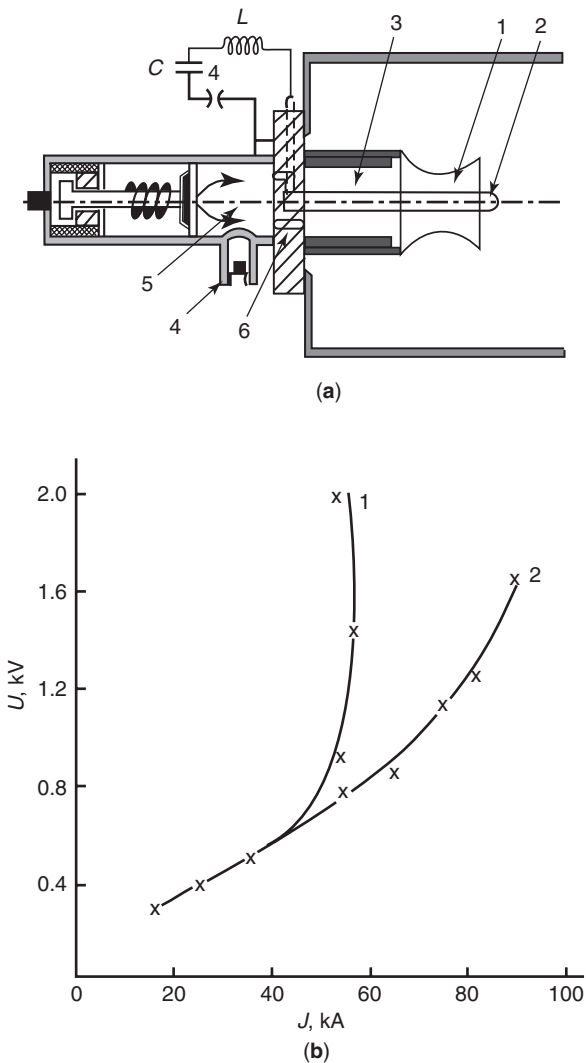


Figure 14. (a) Diagram of a model of a high-power coaxial accelerator for study in a quasi-stationary mode: (1) external electrode; (2) internal electrode; (3) buffer volume; (4) piezosensors; (5) antechamber; (6) diaphragm with openings. (b) Typical volt-ampere curve using hydrogen (1) $m = 1.5 \text{ g/s}$; (2) $m = 4.5 \text{ g/s}$.

cathode, and the exterior wall would serve as the anode after connection to a source of strong current.

Given high conductivity (or more accurately a high magnetic Reynolds number), the flow of plasma in such a system, under the influence of magnetic field pressure created by current flowing through the central electrode and the plasma itself, in elementary theory maximally approximates a gas dynamic system. But this is the case only if the current between the electrodes flows quasi-radially. Here the critical velocity, instead of the usual speed of sound (S_T), is the

so-called speed of fast magnetic sound:

$$C_s = \sqrt{C_T^2 + C_A^2}, \quad C_A = \frac{B}{\sqrt{\mu\rho}}. \quad (6)$$

Here ρ is the plasma density and C_A is the so-called Alven velocity. Evaluation of the exhaust velocity yields a value ~ 100 km/s, given reasonable fields and densities.

However, experiments in the early 1960s at the AEI with such accelerators, in a quasi-stationary mode with discharge currents of 30–100 kA, showed that current at high velocities “does not want” to go along the radius, but starts to slip along the anode, escapes to the edge of the nozzle or even its reverse side, and induces strong erosion of the anode, rather than the cathode, as had originally been assumed. This causes the voltage to increase sharply at discharge (Fig. 14b). Formally, this may be explained by the Hall effect and visually by the dynamics of electrons in mutually perpendicular (E, V) fields. In the accelerator described, the electrons that emerge from the cathode and move toward the anode enter the perpendicular fields. That is, the azimuthal magnetic field and the electric field that (as a result of its equipotentiality) are perpendicular to the anode. As a result, if collisions are infrequent, the electrons begin to drift at velocity U_E not toward the anode, but along it. The presence of a longitudinal current (slippage) along the anode leads, under the influence of Ampere’s force, to squeezing the electrons away from the anode and toward the cathode. This decreases the frequency of collisions and thus fosters an increase in the role of drift. A radical solution of the problems that arise here was found by shifting to ionic current transfer, the creation of anodes able to emit ions, but this is still very difficult.

High-Current End Accelerators HCEA. It became clear that it was necessary somehow to modify the electrodes’ geometry and operate at relatively low discharge current, until the system entered the mode of slipping along the anode. These considerations led to the idea of “end high-current thrusters” proposed by A.A. Porotnikov (Scientific Research Institute for Thermal Processes). The initial idea proposed cutting off the extended electrodes and leaving only the “ends.” This is how the system got its name. But later, during the process of optimization, it became clear that the cathode really did need to be truncated, but that the anodes should be left in the shape of a relatively long nozzle.

The Design and Integral Characteristics of HCEA. An optimized HCEA design is shown in Fig. 15a. There are three main units:

1. a contoured anode
2. a cathode with many rods (wires) connected to the delivery system for the working medium
3. the insulator

The use of a multicavity cathode made of tungsten wire makes it possible to decrease the temperature of this unit significantly and at the same time, use it as

a vaporizer if the working medium is delivered in its liquid phase. This was a real possibility because plans called for using light lithium.

The basic formula underlying the integral performance characteristics of the HCEA is the formula for the thrust developed by the thruster:

$$F_H = \Theta 2 \times 10^{-7} J_A^2. \quad (7)$$

Here Θ is a geometric factor close to 1. From this, it follows that for discharge current $J = 10^4 \text{ A}$, the thrust is of the order of 2 kg.

Crisis Current. The main theoretical problem the developers of the HCEA encountered was obtaining high exhaust velocities by increasing the discharge current with constant expenditure of mass. It turned out that a maximum discharge current J^* exists, which is a function of design characteristics of the accelerator and loss of mass; upon approaching it, the voltage begins to increase sharply despite only small changes in the discharge current (Fig. 15b). This is accompanied by intense fluctuations, and the current begins to slip along the anode and short circuits at the edge of the nozzle or even on the back side of the nozzle. This set of phenomena was named "crisis current." Evidently, this repeats what was observed in the coaxial accelerator described

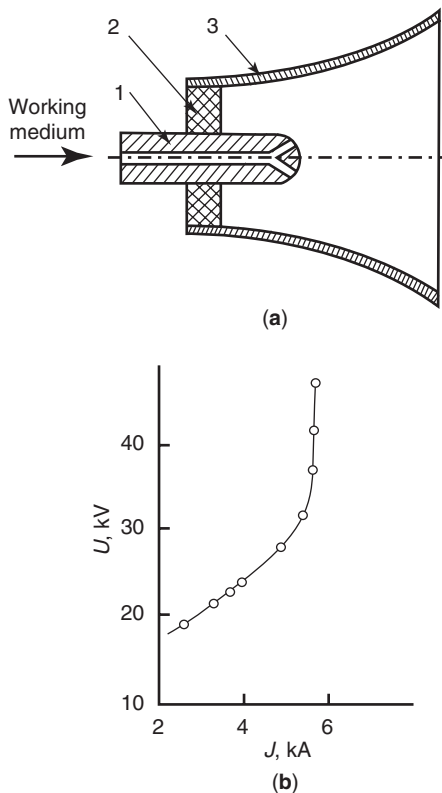
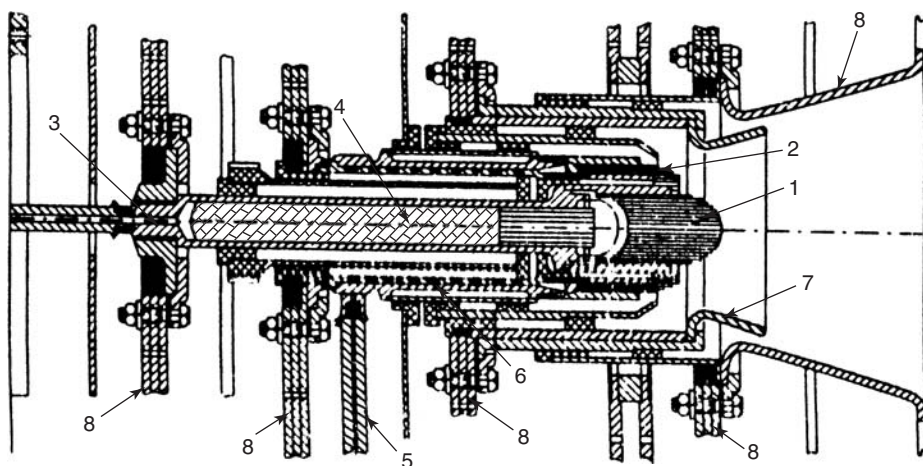
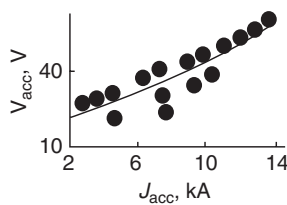


Figure 15. (a) Diagram of the high-power outlet accelerator: (1) cathode; (2) insulator; (3) anode. (b) Typical volt-ampere curve of the high-power outlet accelerator.



(a)



(b)

Figure 16. (a) Test stand version of the high-power two-stage HCEA: (1) central cathode; (2) peripheral cathode; (3, 5) lithium delivery pipes; (4, 6) cavities for generating weakly ionized lithium vapor; (7, 8) first and second stage anodes. (b) Typical volt-ampere curves for the two-stage HCEA.

before. Entrainment of the current at the edge of the nozzle leads to the development of connections and the destruction of the nozzle even when it is made of tungsten.

A large number of experiments established that the critical current is a function of mass expenditure and geometry. The square of critical current proved to be proportional to mass loss. The search for a way to overcome crisis current continues. In particular, the Machine Building Central Institute has developed a two-phase HCEA, where the crisis current is either suppressed or significantly avoided. The design of this HCEA and its volt-ampere curve are shown in Fig. 16. It seems that in this model, ionic current transfer has begun to occur.

Magnetoplasma Dynamic Thrusters (MPDT). The seriousness of the crisis current and also the high level of power needed to maintain comparatively high performance parameters of the HCEA have naturally stimulated a search for modifications that would give rise to high performance at significantly lower power.

It was possible to take the first step by using a simple technique, placing a magnetic field coil on the HCEA (Fig. 17). It should be noted that, independently

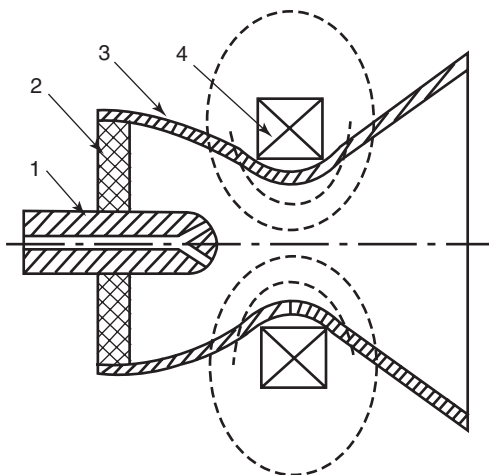


Figure 17. Diagram of the magnetoplasma dynamic thruster: (1) cathode; (2) dielectric; (3) anode; (4) magnetic field coil.

of work in the USSR, a similar design was proposed and studied in the United States and then in Germany and Japan.

As a result, a unique hybrid of the SPT and HCEA was developed. This accelerator actually did increase the performance characteristics in the area of lower power (\sim several tens of kW) compared to the HCEA but did not bring them to the level of the SPT, in spite of the fact that these accelerators were the object of virtually continuous work (although of varying intensity), starting in the early 1950s and continuing to the present.

The physical process in the MPDT proved to be rather complex. Its performance characteristics are sensitive to the magnitude of the discharge current, as well as to the magnitude of the magnetic field and system geometry. For an approximate evaluation of thrust, we may use the formula,

$$F = \frac{J_p B_A D_A}{2} \left(1 - \frac{B_A}{B_K} \right), \quad (8)$$

where B_A and B_K are the induction of the external magnetic field at the ends of the anode and cathode and D_A is the diameter of the anode. The characteristic value is ~ 0.3 – 0.4 .

It proved to be that the MPDT was also subject to the effect of crisis current and the high performance characteristics could be obtained only with lithium at relatively high discharge power. Thus, at the Moscow Aviation Institute, V.B. Tikhonov's group, measured a specific impulse of $I_{sp} \sim 4500$ s and efficiency $\eta \sim 45\%$, given power of $N \sim 150$ kW ($J_p \sim 2000$ A, $U_p \sim 75$ V, and field $B_K = 0.09$ T). This accelerator is depicted in Fig. 18. At JPL in the United States, $I_{sp} \sim 4000$ – 5000 s and efficiency $\eta \sim 40$ – 60% were obtained, given power of ~ 200 kW.



Figure 18. Photograph of a MPDT with power of up to 200 kW. The anode diameter at the end is 160 mm. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

BIBLIOGRAPHY

1. Morozov, A.I. Plasmodynamics. In V.E. Fortov (ed.), *Encyclopedia of Low Temperature Plasma*. Vol. III, Nauka, Moscow, 2000, pp. 383–438.
2. Turner, M.J.L. *Rocket and Space Propulsion*. Springer.
3. Morozov, A.I. *Zhurn. Elect. Techn. Fiz.* 32: 305 (1957).
4. Artsimovich, L.A., et al. *Zhurn. Elect. Techn. Fiz.* 33: 3 (1958).
5. Kozeyev, M.N. Pulsed electronic plasma accelerators. In V.E. Fortov (ed.), *Encyclopedia of Low Temperature Plasma*. Vol. III Nauka, Moscow, 2000.
6. Solbes, F., and R. Vondra. *J. Spacecraft Rockets* 10: 406–410.

7. Micei, M.M., and A.D. Ketsdever (eds). *Prog. Astronaut. and Aeronaut.* 187.
8. Morozov, A.I., and V.V. Savelyev. Fundamentals of stationary plasma thruster theory. In V.V. Kadomtsev and V.D. Shafranov (eds), *Review of Plasma Physics*, Vol. 21. Kluwer Academic/Plenum.
9. Artsimovich, L.A., I.M. Andronov, A.I. Morozov, R.K. Snarskiy, et al. *Kosm. issled.*, 12: 451 (1974).
10. Garner, C., J. Brophy, J. Polk, and L. Pless. Cyclic endurance test of a SPT-100 stationary plasma thruster. *Third Russian-German Conference on Electric Propulsion Engines*, Stuttgart, Germany, July 19–23, 1994.
11. Morozov, A.I., A.I. Bugrova, V.K. Kharchevnikov, et al. *Fiz Pla.* 23: 635 (1998).
12. Bugrova, A.I., N.A. Maslennikov, and A.I. Morozov. *Zh.T.O.*, 61: 45–91 (1991).
13. Yushmanov, Ye.Ye. Radial distribution of potential in a cylindrical magnetic trap.
14. Zharikov, A.V., and YUS Popov. *Zhurn. Tekhn. Fiz.* 37 (1967).
15. Grishin, S.D. Ion and plasma jet engines. In V.E. Fortov (ed.), *Encyclopedia of Low Temperature Plasma*. Nauka, Moscow, 2000, pp. 291–331.
16. Lyapin, Ye.A., and A.V. Seminkin. Current composition of research on anode layer accelerators. In *Ion Injectors and Plasma Accelerators*. Energoatomizdat, Moscow, 1990, pp. 20–33.
17. Garner, C., J. Brophy, J. Polk, S. Seminkin, V. Garkusha, S. Tverdokhlebov, and C. Maresse. Experimental evaluation of Russian anode layer thrusters. *AIAA-94-3010*.
18. Morozov, A.I., U.S. Pavlichenko, V.I. Tereshin, et al. *Plasma Devices and Operation*, 2: 155 (1992).
19. Tichonov, V.B., S.A. Siminichin, V.A. Alexandrov, G.A. Popov, et al. Research of plasma accelerator process in applied magnetic field thrusters, Abstract # 93076, *XXVIII Int. Electric Propulsion Conf.*, Seattle, WA, Sept., 1993.
20. Tichonov, V.B., S.A. Siminichin, J.P. Brophy, and J.E. Polk. Performance of the 130 kV MPD thrusters with external magnetic field and lithium as a propellant, Abstract # 97117, *XVth Int. Electric Propulsion Conf.*, Cleveland, OH, 1997.
21. Polk, J.E., and T.J. Puirotto. Alkali metal for MPD thrusters. Abstract #913572, *AIAA Conf. Adv. SE Technol.*, Cleveland, OH, 1991.
22. Popov, G., N. Antropov, et al. Experimental study of plasma parameters in high efficiency pulsed plasma thrusters. *27 IEPD*, Pasadena, CA, October 15–19, 2001.

ALEXEY I. MOROZOV
Kurchatov Institute
Russian Science Center
Russia

VYACHESLAV M. BALEBANOV
Institute of Space Research
Russian Academy of Sciences
Russia

PLUTO AND CHARON

Pluto and Charon: The Big Picture

Pluto is a small, cold planet on the outskirts of the known solar system. Pluto's diameter (~ 2370 km) is roughly equivalent to the distance from New York to

Las Vegas, or about two-thirds the size of our Moon. Pluto has one known moon of its own, Charon, whose diameter (~ 1250 km) is slightly more than half of Pluto's. Unlike seven of the other eight planets (Uranus is the exception), Pluto rotates on its side, as does Charon in its orbit around Pluto. Charon's distance from Pluto is 19,636 km, compared to the Earth–Moon distance of 384,400 km.

Despite their proximity, Pluto and Charon are covered by bright frosts of differing compositions: H_2O ice covers Charon (1), whereas Pluto's surface is predominantly N_2 frost that has traces of methane and carbon monoxide ices (2). Both objects have densities near 2 gm/cm^3 (3), implying a roughly even mixture of rock and H_2O ice for their bulk compositions.

The Pluto/Charon system has a highly elliptical orbit around the Sun. At the time of Pluto's perihelion in 1989, the system was less than 30 astronomical units (AU) from the sun, but the aphelion separation will be 50 AU in 2123. As Pluto recedes from the Sun, much of its thin N_2 atmosphere will condense as frost on the surface. This periodic redeposition of fresh frost takes place every Pluto year (248 Earth-years) and is the reason that Pluto's surface has one of the highest albedos in the solar system (4).

The Pluto/Charon system currently represents one of the wide open frontiers in the solar system. No spacecraft has visited Pluto/Charon (although NASA has plans for an upcoming Pluto–Kuiper Belt mission). The angular diameter of Pluto is only a tenth of an arcsecond at best, which is just about the resolving limit of the Hubble Space Telescope (HST) and ground-based observatories that have adaptive optics capabilities. Consequently much of what is known about the Pluto/Charon system has been pieced together from indirect clues or observations taken during unique observing geometries.

Observational Milestones: A Chronological Overview

Pluto's discovery was due in large part to the efforts of Percival Lowell, who believed that a "Planet X" must exist to account for so-called perturbations in the orbit of Uranus beyond those caused by the presence of Neptune. In point of fact, Pluto is much too small to cause perceptible perturbations in the orbits of Uranus or Neptune, and the residual "perturbations" disappear when a more accurate mass of Neptune is used to calculate the Uranian orbit. Nevertheless, the prospect of finding a ninth planet motivated the astronomers of Lowell Observatory to mount three separate observational surveys, beginning in 1905 and leading to Clyde Tombaugh's discovery of Pluto in 1930 (Fig. 1; (5)).

In the year after the discovery of Pluto, teams from around the world calculated the size and shape of Pluto's orbit (Fig. 2). As the baseline of observations grew, it became apparent that Pluto's orbit was unusual; its eccentricity was higher (0.25), its inclination greater (17 degrees), and its semimajor axis larger (39.44 AU) than that of any other planet.

Pluto's bulk properties (e.g., size, rotational period, and surface composition) were more difficult to determine primarily because its magnitude (~ 13.3 at the time of discovery) was beyond the capabilities of early photometers and spectrometers. Rotational variations in Pluto's brightness were first observed in the 1950s and led to the important conclusions that Pluto rotated every



Figure 1. Clyde Tombaugh, the discoverer of Pluto, shown here at Lowell Observatory in Flagstaff, Arizona, entering a telescope dome with photographic plates in holders. The background shows two of the discovery blink plates. Pluto is much easier to find with the arrows, which unfortunately were not on the original plates (photos courtesy of Lowell Observatory). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

6.39 days, and Pluto's surface had distinct bright and dark regions that caused a 30% change in Pluto's brightness as they rotated in and out of view. Because Pluto's surface brightness was not known, its size could not be estimated from its observed magnitude. Early estimates of Pluto's size assumed a dark surface (e.g., a 4% reflectance) and therefore vastly overestimated Pluto's size and mass.

In 1976, the first compositional identification was made; a crude spectrum of Pluto revealed the presence of methane (but not ammonia, H_2O , or the other fully hydrogenated compounds of the common nonnoble elements C, N, or O) (6). The presence of methane ice suggested for the first time that Pluto was much brighter (and therefore much smaller) than previously suspected.

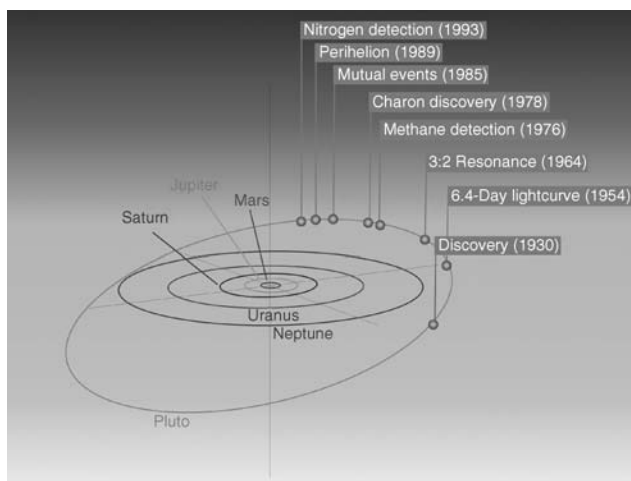


Figure 2. Pluto has traversed only about a third of its orbit since its discovery in 1930. The pace of discoveries has increased since the detection of methane frost on Pluto in 1976 and the discovery of Pluto's satellite, Charon, in 1978. Some of the milestone discoveries are shown here. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

In 1978, Pluto's satellite Charon was discovered from photographic plates showing "bulges" on the side of Pluto's disk (7). A host of information followed from the regular series of eclipses of Pluto by Charon (and vice versa) (see section Mutual Events). The mutual events lasted from 1985 through 1990 and led to the first accurate determinations of the radii, masses, and densities of Pluto and Charon. Eventually, the first maps of the surfaces of Pluto and Charon were constructed from the mutual event light curves. The HST's faint object camera (FOC) was used to image Pluto's disk directly in the mid-1990s, albeit at a slightly lower spatial resolution (8).

In the 1980s, Pluto was observed by the infrared satellite IRAS at wavelengths that ranged from 25–100 μm . In the early 1990s, Pluto was observed at longer wavelengths (0.8 and 1.3 mm) from various ground-based submillimeter telescopes. In theory, each of these long wavelength observations should pin down Pluto's blackbody temperature, but in practice, the IRAS fluxes indicate a much warmer surface than the submillimeter observations. A temperature-dependent spectral feature (see section Spectroscopy) is a third clue to Pluto's surface temperature. In 1997, Pluto was observed by the ISO satellite (a European successor to IRAS), which confirmed that Pluto's thermal emission at 60 and 100 μm is higher than one would expect, given the low submillimeter fluxes. Suffice it to say that Pluto's temperature is currently controversial (see section Surfaces).

In 1988, Pluto was observed as it passed in front of a star (see section Stellar Occultation), an event which resulted in the discovery of Pluto's atmosphere and determinations of Pluto's vertical temperature and pressure profiles (9). Advances in spectroscopic instrumentation led to the discoveries of N_2 , CO , and CH_4 ices on Pluto and H_2O ice on Charon. Directly observed spectra of Charon without Pluto have been obtained only since 1998, using the spatial resolving capabilities of the HST's NICMOS spectrometer or ground-based instruments using adaptive optics.

Orbital Dynamics. Pluto has the largest inclination and eccentricity of all of the nine planets (see Table 1). Pluto's distance from the Sun ranges from 30 to 50 AU, and its vertical distance from the ecliptic plane ranges from $\sim +9$ AU to ~ -14 AU. This large variation has significant implications for Pluto's surface and atmosphere.

For 20 years of its 248-year period, Pluto is closer to the Sun than Neptune, most recently between 21 January 1979 and 14 March 1999. Although Pluto's and Neptune's orbits appear to cross in a bird's-eye view of the solar system, Pluto never comes closer than 18 AU to Neptune. Whenever Pluto is "inside" Neptune's orbit, it is also several AU above Neptune's orbit. More importantly, whenever Pluto is inside Neptune's orbit, Neptune is always at least 70° away from Pluto in its own orbit. This special relationship between Pluto and Neptune is a manifestation of their 3:2 resonance (2 Pluto years are equivalent to 3 Neptune years).

The 3:2 resonance protects Pluto from close encounters with Neptune. If Neptune were to "catch up" to Pluto (such that a collision seemed imminent), Pluto would initially lose angular momentum to Neptune. Pluto would then fall inward toward the Sun and in the process, speed up and pull away from Neptune (smaller heliocentric orbits have faster tangential velocities). It may seem counterintuitive that a loss in angular momentum results in a gain in tangential

Table 1. **Summary**^a

	Pluto	Charon
Bulk properties		
Radius, km	1145–1200	600–650
Mass, 10^{25} g	1.315 ± 0.003	0.156 ± 0.003
Density, g cm^{-3}	1.92–2.06	1.51–1.81
Surface g, m s^{-2}	0.66	0.29
Surface comp.	H ₂ O, N ₂ , CH ₄ , CO	H ₂ O
Atmospheric comp.	N ₂ , CH ₄ , CO(?)	?
Pressure scale ht, km	55.7 ± 4.5	—
Temperature K	35–60	?
Blue albedo	0.44–0.66	0.38
B–V magnitude	0.867 ± 0.008	0.700 ± 0.010
Orbital parameters		
(Heliocentric parameters for Pluto, Plutocentric parameters for Charon)		
Semimajor axis	39.5447 AU	19636 ± 8 km
Eccentricity	0.249050	0.0076 ± 0.0005
Inclination, deg	17.14217	96.163 ± 0.032
Ascending node, deg	110.29714	222.993 ± 0.024
Long. of periapses, deg	224.12486	219.1 ± 0.9
Mean Longitude, deg	238.74394	32.875 ± 0.023
Epoch	JDT 2451545.0	JDT 2449000.5
Period	248.0208 yr	6.387223 ± 0.000017 d
Obliquity, deg	119.6 ± 0.6	—

^aFrom Reference 3, p. 195.

velocity, but examples of this resonant mechanism abound throughout the solar system (e.g., the coorbital satellites of Saturn). When Pluto “catches up” to Neptune, the momentum exchange works in reverse.

Charon’s orbit around Pluto is nearly circular. Both objects continuously present the same face to each other (as our Moon does to Earth), a state called mutually synchronous rotation. In other words, a Charon day, a Pluto day, and a Charon month are all 6.39 Earth-days long. Pluto’s obliquity is close to 120° (compared to 23.3° for Earth), which means that Pluto’s spin axis lies fairly close to its orbital plane.

Masses and Radii. The mass of the Pluto/Charon binary is $1.471 \pm 0.002 \times 10^{25}$ g, as determined from the period and semimajor axis of Charon’s orbit around Pluto. This is 454 times less massive than that of Earth, or more than a million times less massive than Jupiter. The individual masses of Pluto and Charon are less well known, but the current estimates are 1.315 ± 0.003 and $0.156 \pm 0.003 \times 10^{25}$ g. These can be combined with estimates of the radii to get the densities of these bodies. The best measurements for the radii come from the mutual events (see section Mutual Events), and give radii ranging from 1145–1200 km for Pluto and from 600–650 km for Charon. These radii are also consistent with stellar occultations (see section Stellar Occultation), but for one

caveat. The possibility exists that the mutual events are detecting the altitude of a haze layer instead of the surface, in which case Pluto's true surface could lie well below the observed radius. If we ignore this possibility for the moment, we find that Pluto's mean density is $1.92\text{--}2.06\text{ g/cm}^3$ and Charon's is $1.51\text{--}1.81$.

Pluto and Charon have densities roughly halfway between that of water ice ($\sim 1\text{ g/cm}^3$) and rock ($\sim 3\text{ g/cm}^3$). Whether Pluto's interior is differentiated or homogeneous depends on Pluto's thermal history. Differentiation is a runaway process; once started, it generates heat that allows rock to migrate through ice to the planet's core, ending with a rocky core and icy mantle. Current models of Pluto predict a differentiated rocky core comprising 53–71% of the total mass (10), although an undifferentiated Pluto cannot be ruled out at this time. The uncertainties in Charon's density are large enough that we cannot say at this point whether or not Charon's interior is undifferentiated.

Surfaces. Pluto has one of the brightest and most variegated surfaces in the solar system. Two techniques are used to map Pluto's surface: direct imaging using HST's faint object camera (FOC) and an indirect technique based on observations taken during transits of Pluto by Charon (see section Mutual Events). Both techniques show some extremely bright regions (predominantly N_2 frost), often adjacent to very dark regions (composition unknown) (Fig. 3).

Some of the most noticeable features of the mutual event map are a large bright frost-covered region over the southern latitudes (ranging from the pole to around 45°S), an adjacent dark region just south of the equator, a very bright patch at 17°N , and a North Polar region that has an albedo near Pluto's average of 0.5, much darker than the typical southern albedo.

Because Pluto's angular diameter is about 0.11 arcsec (equivalent to a 25-cent piece viewed from a distance of 25 miles), its disk is a challenging target for the FOC. The point-spread function of the FOC has a full width at half maximum (FWHM) that is roughly 0.03 arcsec, only one-third to one-quarter of Pluto's diameter. It is possible to improve on the instrument's spatial resolution by combining FOC images taken at several rotational phases. Unlike the mutual event maps, the FOC map covers all longitudes of Pluto's surface, not just the sub-Charon hemisphere. A comparison of the two maps confirms the presence of a bright South Polar feature, although the mutual event map shows a much larger southern frost-covered region. Both maps show a dark latitudinal band south of the equator, but the bright feature at 17°N in the mutual events map seems to be spread into a wider bright feature in the northern latitudes of the FOC map.

Parts of Pluto are extremely bright, suggesting a layer of recently deposited frost. The bright regions are almost certainly composed of N_2 frost and traces of CH_4 and CO ice (see section Spectroscopy). One of the continuing mysteries, however, is why certain areas are bright and others are not. Pluto's South Pole was in continuous sunlight for most of the past century, yet it remains the largest frost-covered region on the planet. In contrast, the North Pole has been in continuous shadow and ought to be the site where atmospheric N_2 would preferentially condense onto the surface. Surprisingly, it is not nearly as bright as the southern part of the planet. There is an intimate relationship between the atmosphere and the albedo distribution of the surface that is not yet understood (see section Volatile Transport).

Pluto's average surface temperature results from the balance between absorbed sunlight and radiational cooling. On local scales, the surface temper-

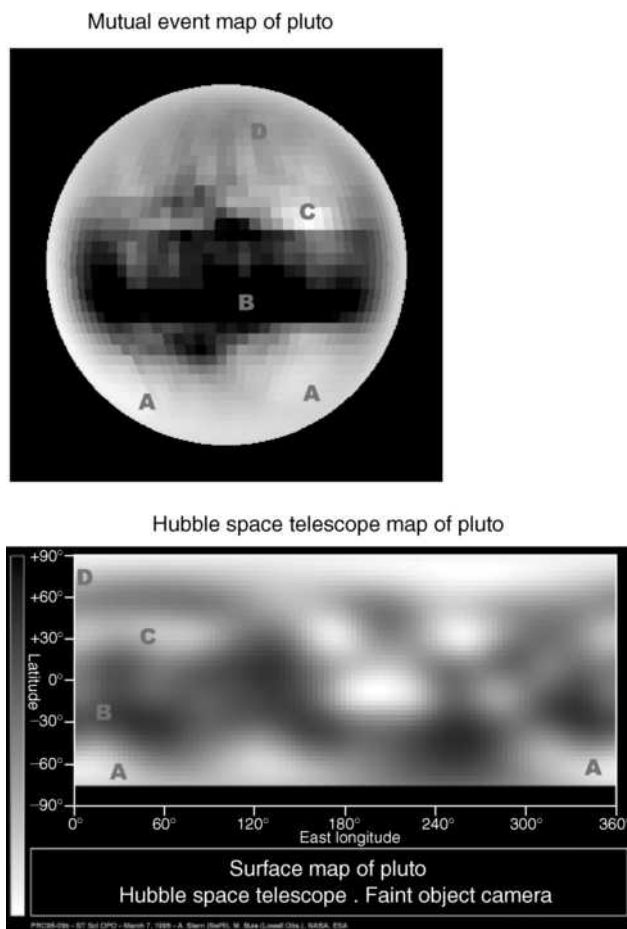


Figure 3. Pluto is a challenging target, even for the Hubble Space Telescope (HST). The highest resolution maps are derived from mutual event light curves (19), but they cover only the part of Pluto's disk that is always below Charon. To compare the mutual event and the HST maps, bear in mind that the center of the mutual event disk is the line of zero longitude, or the left-hand edge of the HST map. Both maps show the presence of a bright southern frost region (A) bordered by a dark band (B). There is a very bright patch at around 17° N (C). Surprisingly, the Northern Hemisphere (D) is only slightly brighter than the average albedo, even though this region should have been a condensation site during the past few decades. (The HST map courtesy of M. Buie, Lowell Observatory). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

atures are also affected by ongoing condensation or sublimation (see section Volatile Transport) in addition to variations in surface brightness and thermal coupling to subsurface layers. It is currently thought that the N_2 frost temperature is 35.6 K across the entire sphere (11), but regions that have less volatile veneers may reach higher temperatures.

Atmospheres. Pluto's atmosphere is composed primarily of N_2 and trace amounts of CH_4 , CO, and photochemical by-products. Because N_2 can exist in both gaseous and solid phases near Pluto's cold surface, the atmosphere is expected to be in vapor pressure equilibrium with the surface. Vapor pressure

equilibrium means that the surface pressure is a function of surface temperature (an extremely steep function in the case of N_2 , CO, and CH_4 ; see Volatile Transport). A drop of only 4 K (from 40 K to 36 K, for example) results in a 10-fold drop in pressure (from 58 μ bar to 5.4 μ bar) (12).

Of the three species seen in the surface spectra, N_2 , CH_4 , and CO, N_2 is by far the most volatile and should therefore dominate the atmospheric composition. CH_4 and CO should also be present in the atmosphere in trace amounts. At the cold temperatures of the outer solar system, subsurface H_2O is as volatile as a rock in your backyard, and no gaseous H_2O is expected to be present in the atmosphere. Gaseous CH_4 has been detected directly in near-infrared spectra and probably forms about 0.1% of the mass of the atmosphere. The vertical distribution of CH_4 is unknown; a few hundred km above Pluto's surface, CH_4 may be depleted due to photochemical destruction, whereas at high altitudes, its relative abundance should rise again as CH_4 diffuses through the heavier N_2 gas. The fraction of CO should be relatively constant with height because it is less strongly involved in photochemical reactions and because the molecular weight of CO is equal to that of N_2 . To date, only an upper limit of 7% has been placed on the abundance of CO in Pluto's atmosphere. However, we can estimate that CO is present in trace amounts (perhaps near $\sim 0.04\%$) from the amount of CO seen in the surface.

The thermal structure of Pluto's atmosphere has been detected directly from about 0.1 to 2.3 μ bar (around 1410 to 1215 km in radius) (see section Stellar Occultation). In this region, the atmosphere is nearly isothermal and has a temperature of 102 ± 9 K. At 1215 ± 11 km, (2.33 ± 0.24 μ bar), the atmosphere undergoes a discontinuity. Below this level, either the atmosphere contains absorbing hazes, or the temperature dives sharply, at roughly 10 K/km, to reach the colder surface temperatures below.

The chemistry of Pluto's atmosphere is highly uncertain due to the unknown ratios of the trace molecules. We can look to the atmospheres of Triton and Titan as a guide because they also consist primarily of N_2 and trace amounts of CH_4 .

The energetic UV photons from the sun are expected to dissociate gaseous N_2 and CH_4 , which will recombine to form heavier hydrocarbons (C_2H_2 , C_2H_6 , C_2H_4 , C_4H_2) and nitriles (HCN , C_2H_3CN , C_4N_2), as well as light by-products [H , H_2 , C , C_2 , N , NH , NH_2D]. Most of the hydrocarbons and nitriles are expected to condense, forming a thin smog in Pluto's atmosphere and eventually settling through the tenuous atmosphere to the surface. To date, the photochemical products have not been reliably detected. The rate of predicted haze production is too low to explain the discontinuity in the stellar occultation light curve. The lighter species and some of the lighter hydrocarbons will diffuse upward and either escape or participate in ionospheric chemistry. Models of the ionosphere predict a peak in the ion density between 1 nbar and 10 pbar, located at roughly one Pluto radius above the surface, that is composed of the ions N^+ , C^+ , N_2^+ , CO^+ , $C_2H_5^+$, and CH_2^+ (13). The identity of the major ion depends on the composition of the neutral atmosphere. For example, if CO is essentially absent, then the ionosphere will be primarily N^+ .

Even trace amounts of CO and CH_4 in the atmosphere have a large effect on the temperature of the atmosphere. CO acts as a universal coolant by absorbing

thermal energy through collisions and reradiating it away (rotational cooling). In contrast, CH_4 absorbs radiation at 2.3 and 3.3 μm and reemits at 7.7 μm . These absorption and emission rates are temperature sensitive and balance at a temperature of 106 K (14). The net result is that CH_4 tries to act as a thermostat that pegs the temperature of the middle atmosphere to ~ 100 K.

Origin Scenarios

The Pluto/Charon binary is such an unusual system that one wonders how it came to be. Specifically, where in the solar system did Pluto form? How did Pluto become a binary system? And how did the Pluto–Charon binary find itself in the safe haven of the 3:2 resonance? As discussed previously [see Masses and Radii], Pluto’s bulk density constrains the fraction of ice and rock that composes Pluto’s interior. This, combined with cosmochemical models of the composition of planetesimals in the early solar system, implies that Pluto formed in the outer solar system, probably past 10 AU.

Pluto is currently in a stable resonance with Neptune. Running time backwards, we see that a hard niche to leave is also a hard niche to enter. There are two scenarios for the insertion of Pluto into the 3:2 resonance. Either Neptune moved outward to capture Pluto, or Pluto experienced one or more damping events, perhaps as collisions with a then-larger population of similar small bodies.

More than 100 small objects in Pluto-like orbits have been discovered since 1992. From a strictly dynamical standpoint, Pluto and Charon are the two largest known members of the Edgeworth–Kuiper Belt (EKB). It is not yet known whether Pluto and Charon were formed by the same processes as the rest of the EKB, or whether they formed elsewhere in the solar system.

The current standard model for the formation of the Pluto/Charon binary is by a giant impact, which could also explain Pluto’s large obliquity. The giant impact scenario supposes that a proto-Pluto was impacted by a large interloper. The scattered debris rapidly reaccreted into the present-day Pluto and Charon, albeit the orbit for Charon was far different. The circularization of Charon’s orbit and the evolution of the mutually synchronous spin periods are expected to take place on a ~ 100 -million-year timescale as a result of tidal dissipation. The alternative formation hypotheses generally have one or more fatal flaws, such as: **Fission of a Rapidly Spinning Molten Body.** It is unlikely that a proto-Pluto was spun up by the accretion of small impacts to the point that it split into a Pluto–Charon pair. Among other considerations, if the spin-up process were not extremely rapid (i.e., not a single large impact!), then the rapidly rotating proto-Pluto would tend to damp out any bar-shaped perturbations to its shape and return to an axially symmetrical state.

Intact Capture of Charon by Pluto. This hypothesis requires some short-lived mechanism by which a rogue Charon’s velocity (relative to Pluto) is decreased, such as tidal dissipation, collision with circumplanetary debris, or a collision with a third body. All of these mechanisms are extremely unlikely.

Coformation of Pluto and Charon. Could a proto-Pluto acquire a ring of material (from planetesimals colliding nearby) from which Charon could form?

Unlikely, because it is difficult to create a ring that has the angular momentum seen in the current Pluto/Charon binary.

Mutual Events

Shortly after the discovery of Charon in 1978, it was realized that Charon's orbit was oriented nearly edge-on toward Earth. In 1985, the first transits of Pluto by Charon were observed, and for the next six years, Charon transited in front of Pluto (interspersed with Pluto occultations of Charon) every 6.39 days. Charons' transit path across Pluto migrated from the northern latitudes in 1985 and 1986 to the southern latitudes in 1989 and 1990.

The 6-year series of transits and occultations are collectively called the mutual events. Once the Pluto–Charon separation was measured (initially by high-spatial-resolution speckle interferometry in 1980), the timing of the beginning and end of events provided the first reliable way to measure Pluto's and Charon's radii. (The average density of the Pluto/Charon system was one of the first mutual event results because it is a function of the timing alone and does not depend on their separation). After Charon's orbit around Pluto was determined, the mutual events could be used to build the first maps of Pluto's surface.

Pluto and Charon obscured each other regularly every 6.39 days throughout the mutual event season (1985 through 1990). Although both objects were too small to resolve, it was easy to see the system brightness decrease as part of Pluto was covered up. An event typically lasted from 1–4 hours; during that time, the Pluto/Charon brightness might decrease by as much as 60% and then return to normal levels. If you could keep track of which part of Pluto was covered at any time and you knew the amount of dimming that took place when that part was covered up, you would know the surface brightness of the covered-up region. After observing a number of transits, you could piece together a brightness mosaic of Pluto (Fig. 4). More precisely, you could map half of Pluto by this technique. Just as our Moon always shows only one side to Earth, Pluto and Charon are locked in a mutually synchronous rotation, which means that Charon always transits the same face of Pluto.

Spectroscopy

Much of what is known about Pluto and Charon lies in the details of their infrared spectra. On Pluto, the most obvious constituent is CH_4 , followed by CO and N_2 (2). On Charon, the only certain surface frost is H_2O , although NH_3 has been suggested as a trace constituent (15). Pluto is one of the redder objects in the solar system. It is currently thought that UV radiation photolyzes CH_4 and N_2 into larger hydrocarbons, similar to the processes in smog-producing regions on Earth. These photolytic by-products have been produced in the laboratory but have not yet been identified in the spectra of Pluto or Charon. In the lab, they are initially red in color and become brown or black as they undergo more photolytic processing.

The most prevalent surface frost on Pluto happens to have the weakest spectral signature. The $2.14\text{-}\mu\text{m}$ absorption feature of N_2 is barely detectable in Pluto's spectrum, but the intrinsic band strength is so low that its mere presence

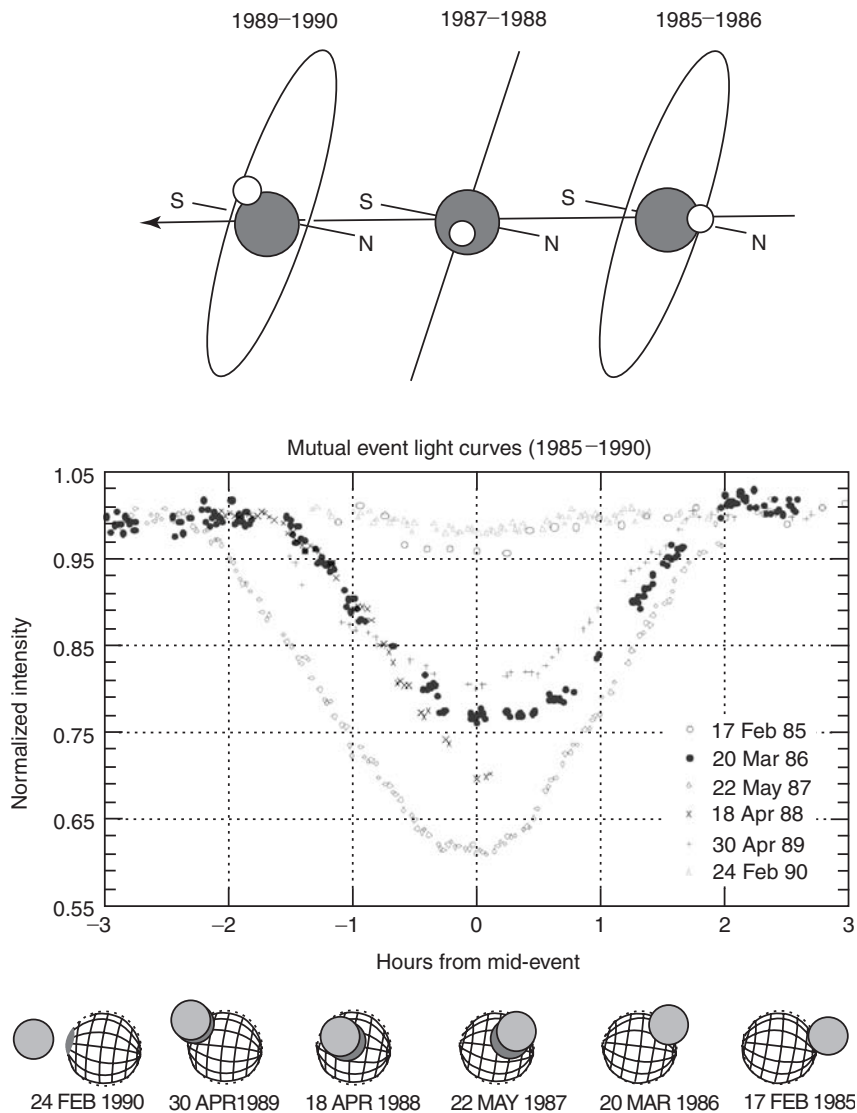


Figure 4. From 1985 through 1990, Pluto and Charon occulted or transited each other nearly weekly, a series that is collectively called “mutual events.” These produced a host of results, including new estimates for the sizes of Charon and Pluto, the orbit of Charon, and even a crude map of Pluto’s surface. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

means that the surface frost on Pluto is at least 95% nitrogen ice. Furthermore, the shape of the N_2 spectral feature changes with temperature, a fact that has been exploited to use Pluto’s spectrum as a thermometer. Lab work performed in the early 1990s suggested that the N_2 absorption feature matched a frost temperature of 40 ± 2 K (16), but more recent laboratory spectroscopy suggests that 36 K is a closer match (11). N_2 has a phase transition at 35.6 K as it changes from a cubic to a hexagonal crystal. It is possible that subsurface N_2 ice undergoing

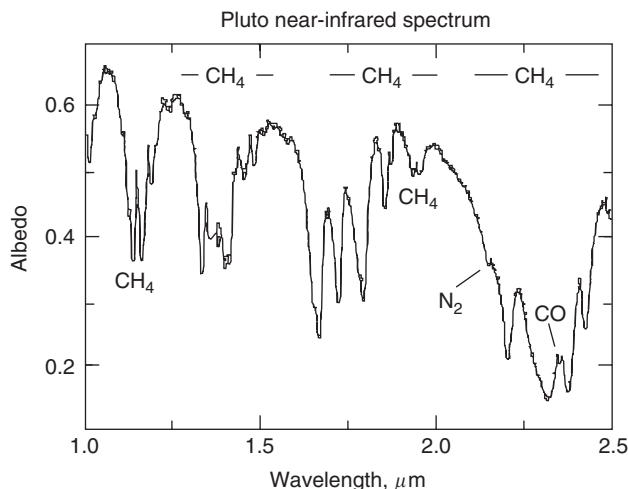


Figure 5. Pluto's infrared spectrum shows the presence of nitrogen, methane, and carbon monoxide frosts.

this phase change acts as an energy storage mechanism and thereby maintains a frost temperature near 35.6 K. It is worth pointing out that a difference of a few degrees in the N_2 frost temperature has an enormous effect on the total mass of Pluto's atmosphere because the N_2 vapor pressure is an extremely steep function of temperature (see Volatile Transport).

CH_4 absorption features dominate Pluto's spectrum. The spectroscopy of Pluto shows two distinct components of CH_4 , one frozen in an N_2 frost matrix and one nearly undiluted (17). CH_4 is about 10,000 times less volatile than N_2 and CO at Pluto's temperatures, so it is conceivable that some separation of volatiles occurs as different molecules condense onto or sublime from the surface under different conditions (Fig. 5).

Since the early mutual events, it has been known that Charon's surface is primarily H_2O ice. Spectra obtained in the past year of Charon without Pluto show that the H_2O ice is crystalline rather than amorphous. This finding is somewhat surprising because irradiation by electrons, fast protons, or UV photons will transform crystalline ice into amorphous ice over time, provided that the ice temperature is lower than 70–80 K. The fact that Charon is predominantly covered with crystalline ice implies that Charon experienced a recent resurfacing event or that Charon's surface somehow reaches temperatures that are high enough to allow the amorphous-to-crystalline transition.

Stellar Occultation

Pluto occasionally passes directly between Earth and a star. These events are called stellar occultations, and they represent rare opportunities to study Pluto's atmosphere. In practice, these events are hard to observe; they require a star that is comparable to Pluto's brightness (e.g., fourteenth magnitude or brighter) and extremely accurate relative positions between Pluto and the star (Fig. 6).

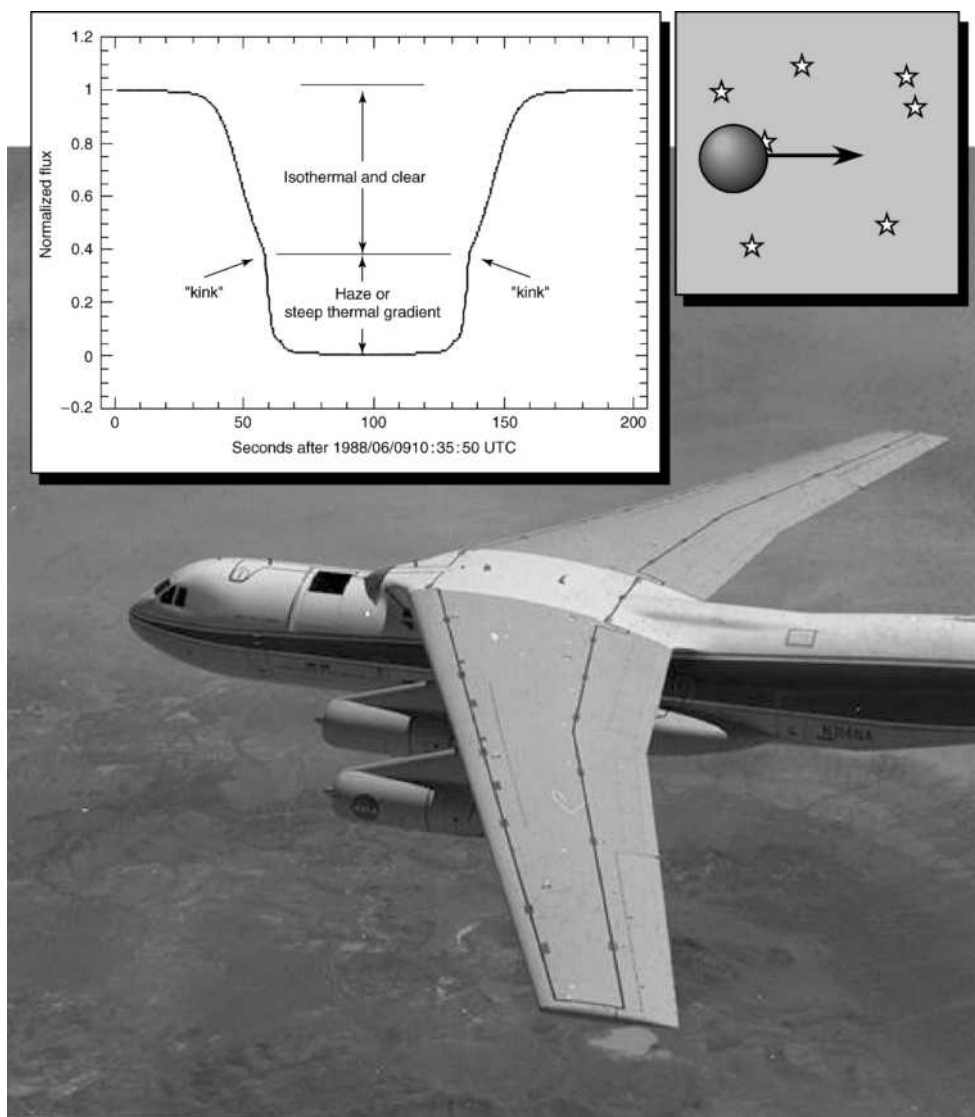


Figure 6. Occasionally, Pluto moves briefly in front of a background star in an event called a “stellar occultation.” One of the best occultation light curves was obtained from NASA’s Kuiper Airborne Observatory, a modified C-141 housing a 1-meter telescope. An unusual feature in Pluto’s occultation light curve is the sudden change in slope (the “kink”) that occurs partway through both the stellar ingress and egress. Possible explanations include a haze layer or a thermal inversion in Pluto’s lower atmosphere. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Pluto casts a fast-moving shadow that is the size of its disk. This shadow typically crosses a ground station in about a minute. The key to observing a stellar occultation is being in the right place at the right time; the shadow path predictions often jump by thousands of km as calibration observations of Pluto and the target star refine the track of the shadow’s path. Often observers find

themselves with airline tickets to exotic places, only to find that the shadow no longer intersects Earth at all! Given these difficulties, it is not surprising that there has been only a single successful observation of a stellar occultation by Pluto to date.

On 8 June, 1988, Pluto moved in front of a twelfth magnitude star. If Pluto had had no atmosphere, the light from the star would have blinked out abruptly as Pluto's disk blocked the star (and would have reappeared equally abruptly moments later). However, the starlight faded gradually. This event marked the discovery of Pluto's atmosphere.

The 1988 occultation provides the most detailed and direct information we have about Pluto's atmosphere. Most of the star's initial disappearance is not due to Pluto's atmosphere blocking the light, but rather due to Pluto's atmosphere acting like a lens that spreads out the light, differential refraction caused by the exponential vertical profile of atmospheric pressure. At the initial stages of the occultation, before the star has been attenuated to the half-light level, the light curve follows a shape consistent with Pluto's atmosphere being clear and nearly isothermal. There is a discontinuity at the half-light point beyond which the rate of attenuation suddenly becomes steeper. The interpretation of this "knee" in the light curve is currently ambiguous; it could be caused by an absorbing haze layer, a steep thermal gradient (which changes the refractive properties of the atmosphere across a range of altitudes), or a combination of the two (18).

Volatile Transport

Although Pluto's atmosphere is currently a meager 40 μ bar, it is sufficient to transport volatiles across Pluto's surface. Pluto's atmosphere and surface frosts form an intimately coupled system, primarily because the active volatile species (N_2 , CO, and CH_4) have vapor pressures that are extremely steep functions of temperature.

We believe that the general transport of volatile materials on Pluto follows a simple rule: frost sublimates from Pluto's sunlit side and condenses on its shaded side. Weather prediction on Pluto ought to be equally simple: the prevailing winds should always blow radially outward from the subsolar point. To see why this must be so, consider a patch of N_2 frost on Pluto's daytime hemisphere. The long day tends to heat up the frost patch, but any rise in temperature, even by a degree or two, would result in a dramatic rise in the local surface pressure. Large pressure gradients cannot be sustained; instead there is a local flow outward from the sunlit region.

A side effect of the steep vapor pressure/temperature relation is that the entire frost-covered areas of the planet should maintain a single global temperature. The sunlit regions are cooled by the latent heat of sublimation of N_2 frost, and the shaded regions are warmed by the latent heat of condensation. A local frost patch cannot grow too warm or too cold; that would imply an unsustainable pressure excess or deficit in that neighborhood. The global frost temperature model breaks down only if Pluto's atmosphere is too rarified to propagate pressure gradients from the dayside to the nightside (as is so around the volcanos of Io, for example). In the rarefied atmosphere, local atmospheres

freeze out on the cold, shaded surface before reaching all parts of the planet, even though the flow emanates at supersonic velocities from its sources. Pluto's atmosphere is currently thick enough to equilibrate frost temperatures across the entire surface, but as Pluto recedes from the Sun in the coming century, its atmosphere may freeze out to the point that only a local, dayside atmosphere remains. If that does occur, the nightside frosts will become significantly colder than the dayside frosts. It is important to note the shortcomings of the simple sublimation-balance model. The most significant problem is that the present frost distribution does not match the model's predictions for frost deposition on the North and South Polar regions. Because Pluto rotates "on its side," the South Pole experienced a long summer (up to 124 years of continuous sunlight) until 1989, yet the South Pole was not devoid of frost. On the contrary, the Southern Hemisphere appears to be the site of Pluto's largest frost reservoir. The predicted depth of new frost in northern latitudes is roughly 1 m during the 124-year northern winter, easily enough to change Pluto's albedo. Yet such an albedo change is not apparent in the mutual event maps. A more detailed frost transport model is necessary that accounts for the numerous influences on the surface temperature, the horizontal transport of sublimated N₂, and the microphysical state (e.g., porosity) of the surface frost.

BIBLIOGRAPHY

1. Buie, M.W., D.P. Cruikshank, L.A. Lebofsky, and E.F. Tedesco. Water frost on Charon. *Nature* 329: 522–523 (1987).
2. Owen, T.C., T.L. Roush, D.P. Cruikshank, J.L. Elliot, L.A. Young, C. de Bergh, B. Schmitt, T.R. Geballe, R.H. Brown, and M.J. Bartholomew. Surface ices and the atmospheric composition of Pluto. *Science* 261 (5122): 745 (1993).
3. Tholen, D.J., and M.W. Buie. Bulk properties of Pluto and Charon. In *Pluto and Charon*, edited by S.A. Stern, D.J. Tholen with the editorial assistance of A.S. Ruskin, M.L. Guerrieri, and M.S. Matthews. University of Arizona Press, Tucson, 1997.
4. Stern, S.A., L.M. Trafton, and G.R. Gladstone. Why is Pluto bright? Implications of the albedo and lightcurve behavior of Pluto. *Icarus* 75: 485–498 (1988).
5. Hoyt, M.H. *Planets X and Pluto*. University of Arizona Press, Tucson, 1980.
6. Cruikshank, D.P., C.B. Pilcher, and D. Morrison. Pluto: Evidence for methane frost. *Science* 194: 835–837 (1976).
7. Christy, J.W., and R.S. Harrington. The satellite of Pluto. *Astron. J.* 83: 1005–1008 (1978).
8. Stern, S.A., M.W. Buie, and L.M. Trafton. HST high-resolution images and maps of Pluto. *Astron. J.* 113: 827 (1997).
9. Elliot, J.L., E.W. Dunham, A.S. Bosh, S.M. Slivan, L.A. Young, L.H. Wasserman, and R.L. Millis. Pluto's atmosphere. *Icarus* 77: 148–170 (1989).
10. McKinnon, W.B., D.P. Simonelli, and G. Schubert. Composition, internal structure, and thermal evolution of Pluto and Charon. In *Pluto and Charon*, edited by S.A. Stern, and D.J. Tholen with the editorial assistance of A.S. Ruskin, M.L. Guerrieri, and M.S. Matthews. University of Arizona Press, Tucson, 1997.
11. Quirico, E., and B. Schmitt. Near infrared spectroscopy of simple hydrocarbons and carbon oxides diluted in solid N₂ and as pure ices: Implications for Triton and Pluto. *Icarus* 127: 354–378 (1999).

12. Brown, G.N., and W.T. Zeigler. Vapor pressure and heats of vaporization and sublimation of liquids and solids of interest in cryogenics below 1-atm pressure. In *Advances in Cryogenic Engineering* 25. Plenum Press New York, 1980.
13. Summers, M.E., D.F. Strobel, and G.R. Gladstone. Chemical models of Pluto's atmosphere. In *Pluto and Charon*, edited by S.A. Stern, and D.J. Tholen with the editorial assistance of A.S. Ruskin, M.L. Guerrieri, and M.S. Matthews. University of Arizona Press, Tucson, 1997.
14. Yelle, R.V., and J.I. Lunine. Evidence for a molecule heavier than methane in the atmosphere of Pluto. *Nature* 339: 288–290 (1989).
15. Brown, M.E., and W.M. Calvin. Evidence for crystalline water and ammonia ices on Pluto's satellite Charon. *Science* 287: 107 (2000).
16. Tryka, K.A., R.H. Brown, D.P. Cruikshank, T.C. Owen, T.R. Geballe, and C. Debergh. Temperature of nitrogen ice on Pluto and its implications for flux measurements. *Icarus* 112 (2): 513–527 (1994).
17. Douté, S., B. Schmitt, E. Quirico, T.C. Owen, D.P. Cruikshank, C. de Bergh, T.R. Geballe, and T.L. Roush. Evidence for methane segregation at the surface of Pluto. *Icarus* 142 (Issue Icarus): 421–444 (1999).
18. Young, L.A. Bulk properties and atmospheric structure of Pluto and Charon. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1994.
19. Young, E.F., K. Galdamez, M.W. Buie, R.P. Binzel, and D.J. Tholen. Mapping the variegated surface of Pluto. *Astron. J.* 117 (2): 1063–1076 (1999).

ELIOT YOUNG
LESLIE YOUNG
Southwest Research Institute
Boulder, Colorado

PRECISION ORBIT DETERMINATION FOR EARTH OBSERVATION SYSTEMS

Introduction

The orbit determination problem has its origin in the early efforts of solar system astronomers attempting to describe the motions of the planets and comets as they orbit the Sun. In these studies, the observations of the bodies, as observed from the surface of Earth, were fit to a path through the heavens that was completely described by six parameters. The foundations of modern estimation theory evolved from the early attempts to develop techniques to determine the six fundamental orbital parameters. Three of the parameters determine the orientation of the orbit or trajectory plane in space, and three locate the body in the orbital plane. These six parameters are uniquely related to the position and velocity of the satellite at a given epoch. Six appropriately selected observations will yield a solution for the trajectory. This is the classic orbit determination problem in which there is a match between the number of observations and the parameters

to be determined. However, when the number of observations exceeds the number of parameters to be assigned, special techniques were required to allow using all observations. One solution to this problem was the method of least squares, which was proposed by Gauss in 1795 before his twentieth birthday. In an independent study, Legendre published a similar method in 1806 that led to considerable early debate about who originated the method. This early activity was followed by a period of intense study, culminating in the current theory for estimating dynamic parameters using observations corrupted by random measurement error.

During the past two decades, the requirements for highly accurate determinations of the orbits of near-Earth satellites have been driven by the evolution of the fields of satellite geodesy and satellite oceanography. The ability to use satellite altimeter measurements to obtain accurate, globally distributed, and temporally dense observations of a satellite height as it traverses the ocean surface has opened a new era in oceanography. The ability to use accurate range and range-rate measurements between an orbiting satellite and tracking systems located on Earth's surface has provided a dramatic improvement in the ability to monitor tectonic surface deformation and subsidence and to monitor small but important changes in Earth's rotation. These same measurements, along with satellite-to-satellite measurements, are providing unparalleled views of Earth's gravity field and the gravitational signals from temporal variations in Earth's mass distribution. The advances in each of these fields is tied to the advances in our ability to determine, at high accuracy, the path followed by an Earth-orbiting satellite. The methodology whereby this task is accomplished is referred to as *precision orbit determination* (POD).

In alternative applications, the recent advances in the spatial resolution of orbiting microwave and multispectral imaging systems has stimulated the requirements for accurate near-real-time orbits. These requirements along with the operational challenges of space object catalog maintenance to support collision avoidance in spacecraft operations has stimulated the need for accurate orbit prediction. The improvements in our knowledge of the models for the forces that influence a satellite's motion along with the dramatic improvement in computational capability has opened a new era in determining and predicting the orbits of near-Earth satellites.

The Orbit Determination Concept

The solution of the orbit determination problem involves four fundamental elements: (1) a set of differential equations that describes the motion of the satellite; (2) a numerical integration procedure to obtain a solution of the differential equations; (3) accurate observations of the satellite's motion; and (4) an appropriate estimation method that combines the results of the first three to yield an estimate of the satellite's position, velocity, and appropriate model parameters (e.g., the drag coefficient). The basic procedure starts with an initial model of the trajectory of a satellite during some time interval. This initial orbit will be incorrect due to errors in the estimate for the starting point, deficiencies in the mathematical model for the forces acting on the satellite, and errors in the parameters used in the model. To correct the model, independent observations of the satellite's motion

must be obtained. These observations generally measure only some component of the motion, such as the distance or rate of change of the distance between the satellite and a ground-based tracking station. Measurements of the full three-dimensional position or velocity are usually not available, but as long as the observations depend on the satellite's motion, they contain information that helps to determine the orbit. The evolution of the satellite's position and velocity must be consistent with both the physics of the mathematical model and the sequence of observations, which constrains the orbit estimate to a specific solution.

The observations must also have a corresponding mathematical model to be usable in the orbit estimation problem. The observation model depends on the satellite's motion, and also on the orientation of the spacecraft and the motion of the observing station. The measurement model must relate the location of the tracking instrument to the spacecraft's center of mass, which may change with time as onboard fuel is consumed. At the same time, the tracking station is on a rotating Earth. The observing "station" may even be another orbiting satellite, such as a *Global Positioning Satellite* (GPS). Finally, the measurement model must account for various atmospheric refractive effects and other instrument effects.

Assuming that the measurements are reliable, the discrepancies between the computed observables and the real observations (called the residuals) contain measurement errors and the effects of errors in the initial conditions as well as deficiencies in the dynamic and observational models. Through a linearized least-squares solution process discussed later, the initial conditions and selected model parameters are adjusted to minimize the residuals. Considerable experience is required in choosing the model parameters that are best suited for adjustment. The mathematical models will always be imperfect in some respects, and the adjusted parameters are chosen for their ability to compensate for the deficiencies. The orbit is then recalculated on the basis of improved initial conditions and parameters, the observations are again compared with their computed counterparts, and the initial conditions and parameters are adjusted again. During this iterative process, unreliable observations can be identified and removed. Given a set of observations that contain sufficient information, the adjustments become smaller and smaller with each iteration, and the process is judged to have converged when a satisfactory and stable orbit solution is obtained.

The advent of the GPS and space-qualified GPS receivers have allowed continuous kinematic (i.e., purely geometric) positioning of satellites. However, the dynamic techniques discussed here still tend to provide the best results.

Dynamics of Satellite Motion

The precise trajectory of a satellite has generally been obtained by integrating Newton's dynamic equations of motion by numerical methods (1). The mathematical representation of the motion of the center of mass of a spacecraft is given by

$$\begin{aligned}\vec{r}(t) &= \vec{r}_0 + \int_0^t \vec{v} dt, \\ \vec{v} &= \vec{v}_0 + \int_0^t \vec{a}(\vec{r}, \vec{v}, t, p) dt,\end{aligned}\tag{1}$$

where t is time; \vec{r} and \vec{v} are the position and velocity vectors of the spacecraft's center of mass whose initial values are \vec{r}_0 and \vec{v}_0 at time t_0 ; \vec{a} is the acceleration (force per unit mass) of the spacecraft; and p represents all of the parameters that are employed in the models for the reference frame, the forces, and the observations. The arc length is the time interval from the initial point to some chosen final time. This may be several hours, a day, several days, or longer.

The forces acting on the satellite can be broadly classified as either gravitational or nongravitational. Among the gravitational forces, the two-body term (where the central body is assumed to be perfectly spherical) dominates the orbital motion by far. As a consequence, an orbit is well characterized by the Keplerian elements of an elliptical orbit (2,3). Figure 1 illustrates the geometric properties of the usual set of orbital elements used to describe the motion of a satellite in orbit about the Earth. The satellite height is characterized by the orbit's semimajor axis a , the variation in the radial distance due to the ellipticity of the orbit (the eccentricity e), and the angular distance v (the true anomaly) from the point of closest approach in the orbit (called the perigee). Other angular measures besides v are used to relate time and the motion of the satellite along the orbit, including the eccentric anomaly E and the mean anomaly M . The tilt and orientation of the orbital plane are given by the inclination i and the longitude of the ascending node Ω . These two angles are related to the out-of-plane components of the orbit. The argument of perigee ω is the angular distance along the orbit from the equatorial plane to the perigee, which determines the orientation of the long axis of the elliptical orbit within the orbital plane. The motion of a satellite in Earth orbit is principally characterized by these six orbital elements; the satellite is moving along an elliptical orbit within a plane that tends to precess slowly in space.

The Earth's gravity field is, however, not perfectly spherical, and undulations in the gravity field, corresponding to the variations in Earth's shape and

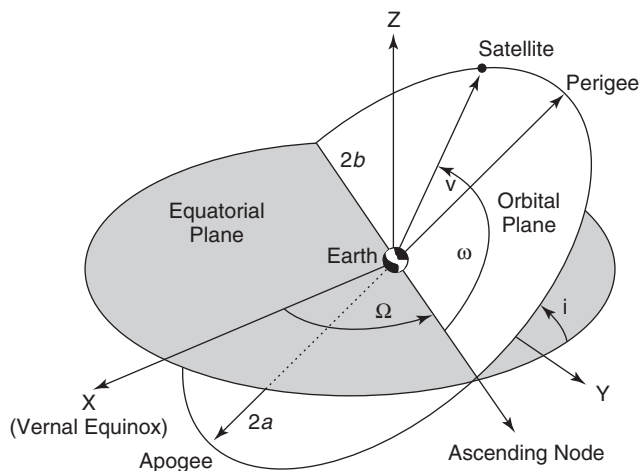


Figure 1. The elements of an elliptical orbit: semimajor axis a , inclination i , longitude (or right ascension) of the ascending node Ω , argument of perigee ω , and true anomaly v . The eccentricity e relates the semimajor axis a to the semiminor axis b through the expression $e^2 = (a^2 - b^2)/a^2$.

density (discussed in other articles), cause perturbations from perfectly elliptical motion (4,5). In addition, perturbations are caused by the gravitational attraction of the Sun, Moon, and planets. For highly precise applications, the effects of general relativity, principally the precession of perigee and the relativistic effects on the observations, must also be considered.

The nongravitational perturbing forces (also known as surface forces) can be broadly classified into the categories of atmospheric drag and radiative pressure (6). Atmospheric drag is typically the greatest concern for artificial satellites because the atmosphere constantly dissipates energy from the orbit. In addition, there are large, sometimes rapid, variations in the atmospheric density at lower altitudes due to geomagnetic activity and solar storms. Radiative pressure includes the effects of direct solar light pressure on a satellite, indirect reflected light and reemitted heat from Earth, and heat emitted by the satellite itself. Finally, the satellite may be subjected to thrust from onboard propulsion or attitude control systems.

Orbit Estimation Problem

As noted earlier, knowledge of the forces acting on a satellite is imperfect, and the initial conditions required to start the integration cannot be known exactly. Observations of the position or velocity of the satellite must be obtained and incorporated into an orbit determination procedure that estimates these initial conditions and corrections to the parameters in the force and measurement models. Starting with estimates for the initial satellite position and velocity, the orbit is predicted for the times of the observations. Using a model for the measurement based on a priori estimates for the tracking station position, Earth orientation, atmospheric refractive effects, and measurement biases, a “computed observation” G is formed and compared to the actual observation Y . The measurement residual $Y - G$ is an indication of the mismatch between the mathematical model and the actual orbit, as illustrated in Fig. 2. Using an appropriate estimation procedure, the residuals from the data fits can be used to

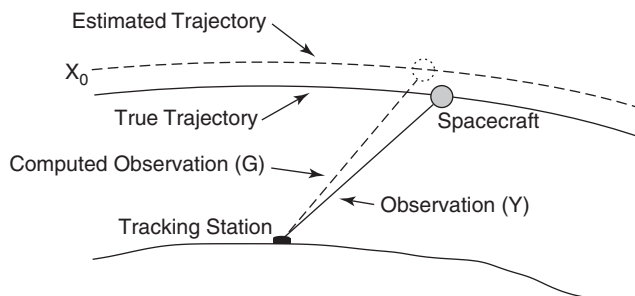


Figure 2. The orbit estimation problem. An observation Y , such as the range to the satellite, is compared to the range G computed from the numerically integrated satellite position to compute a residual $Y - G$. With many such observations, the initial conditions X_0 and other model parameters are iteratively improved to bring the estimated orbit closer to the true orbit by minimizing the residuals in a least-squares sense.

improve the estimates for the initial satellite state, as well as various parameters in the force or measurement model. Clearly, the more observations available and the more accurate they are, the better the orbit can be determined.

This is a complex and nonlinear estimation problem that is usually linearized and solved numerically through an iterative method (7–9). This may be a “batch” process in which all of the measurements within the arc are combined to form the estimates of the initial conditions and the other force and measurement model parameters. Various “sequential” methods are also available in which the orbit and model parameter estimates are updated as soon as new data are available. The fidelity of the models for the dynamics and the measurements, the precision of the tracking data, and the quality of the orbit determination technique determine the accuracy of the resulting satellite ephemerides.

State Equations

As defined, the state vector is composed of all dependent variables or constant parameters required to define the time rate of change of the state of a dynamic system. By this definition, the n -dimensional state vector X can be expressed as follows:

$$X^T = [\zeta^T : \alpha^T], \quad (2)$$

and the state equations become

$$\dot{X} = F(X, t); \quad X(t_0) = X_0. \quad (3)$$

When X_0 in Equation 3 is specified, the motion of the dynamic system will be determined uniquely by the solution to Equation 3. This solution can be expressed formally as

$$X(t) = \Theta(X_0, t_0, t). \quad (4)$$

In the usual orbit determination problem, X_0 will not be known perfectly and, consequently, the true solution $X(t)$ will differ from the nominal solution $X^*(t)$, obtained by using the specified initial state X_0^* . As a consequence, observations of the motion must be made to determine the true motion.

The Observation-State Relationship

Usually, the state vector $X(t)$ cannot be observed directly. Instead, the observations will be a nonlinear function of the state. The observations are influenced by random instrumental errors. As a consequence of these points, the observation-state relationship can be expressed as

$$Y_i = G(X_i, \beta, t_i) + \varepsilon_i, \quad (5)$$

where Y_i is a p -vector of observations at the time epoch t_i , $G(X_i, \beta, t_i)$ is a p -vector of nonlinear functions, β represents the constant parameters that appear in the

observation-state relations, and ε_i is a p -vector of observation errors. As an example, the observation may be the range between a tracking station and the satellite and in this case, Equation 5 would be expressed as

$$R_i = [\bar{\rho} \cdot \bar{\rho}]_i^{1/2} + \varepsilon_i,$$

where $\bar{\rho} = \vec{r} - \vec{r}_S$, \vec{r} is the location of the satellite's laser reflector and \vec{r}_S is the location of the laser ranging station. For this case, Equation 5 will be a scalar observation-state relation at time t_i . The quantity R_i will be obtained as the result of a direct measurement of some physical quantity.

More than one quantity may be observed in a given epoch, that is Y_i may consist of measurements of range, range rate, azimuth, and elevation, where the observed quantities are measured from some observation point on the Earth's surface or from another satellite. The actual physical measurements are physical quantities such as time of flight, phase, or frequency. These measurements can be used to infer range, biased range or range rate, respectively.

If p is the number of observations and n is the number of parameters to be estimated, the following comments can be made regarding the observation-state relationship expressed in Equation 5:

1. At any time epoch, Y_i will be smaller than X_i , that is, $p \leq n$, for the orbit determination problem.
2. Furthermore, even if $p = n$, X_i cannot be correctly determined from Equation 5 unless $\varepsilon_i = 0$, that is, unless the observations are perfect and unless certain observing conditions are met.

To overcome these limitations, a large number of observations, distributed over the arc of interest, are acquired. The information used for estimating the initial state can be expressed as

$$Y_i = G(X_i, t_i) + \varepsilon_i, \quad i = 1, \dots, l. \quad (6)$$

Because $X_i = \Theta(X_0, t_0, t_i)$ from Equation 4, it follows that Equations 6 will represent $l \times p$ equations in $(l \times p \times n)$ unknowns, that is, the unknown components of the observation error and the initial state. Because the number of equations is fewer than the number of unknowns in Equations 6, some additional criterion must be adopted for choosing X_0 . One frequently applied approach is to select the estimate of the state at t_0 , that is, \hat{X}_0 , to minimize the error in the estimate of the initial state. Implementation of this criterion leads to the estimate of X_0 based on either a least-squares error criterion or a statistically based criterion such as the minimum variance or maximum likelihood estimation criterion.

The difficulties of obtaining an estimate of X_0 using Equations 4 and 5 lead to a nonlinear iterative approach based on using linear estimation techniques. This procedure, outlined below, is described in detail by Tapley et al. (10).

Linearization of the Nonlinear Problem

If the true value of the state $X(t)$ and some initially assumed value of the state $X^*(t)$ are sufficiently close throughout some time interval of interest, $t_0 \leq t \leq t_1$, a Taylor series expansion about the initially assumed, or nominal, trajectory at each point can be used to linearize the nonlinear problem. To accomplish this, let

$$\begin{aligned} x(t) &= X(t) - X^*(t); \quad t_0 \leq t \leq t_1 \\ y_i &= Y_i - G(X_i^*, t_i); \quad i = 1, \dots, l. \end{aligned} \quad (7)$$

Then substituting Equations 7 in Equations 3 and 5, expanding in a Taylor's series, and neglecting terms of $O[(X_i - X_i^*)^2]$, the following results are obtained:

$$\begin{aligned} \dot{x} &= A(t)x, \quad x(t_0) = x_0, \\ y_i &= \tilde{H}_i x_i + \varepsilon_i, \quad i = 1, \dots, l. \end{aligned} \quad (8)$$

In Equation 8,

$$A(t) = [\partial F / \partial X]^*, \quad \tilde{H}_i = [\partial G / \partial X]_i^*. \quad (9)$$

The unknown state, X_0 is replaced by the unknown state deviation x_k , and the nonlinear estimation problem is replaced by an estimation problem in which the observations are related linearly to the state and the state is propagated by a system of linear equations that has time-dependent coefficients. The quantity $x(t)$ to be estimated, is the deviation from the reference solution $X^*(t)$. Equations 3 must be integrated using some initial value X_0^* to obtain a reference solution for evaluating $A(t)$ and \tilde{H}_i .

Reduction of Observations to a Single Epoch

To reduce $(l \times n + m)$ unknowns in Equations 8 to $m + n$ unknowns where l is the number of observation epochs, n is the number of parameters in x , and m is the number of observations errors, all of the state variables x_i are expressed as a function of the state at a single epoch, say x_k . To this end, note that the solution of the first of Equations 8 can be expressed as

$$x_i = \Phi(t_i, t_k) x_k, \quad (10)$$

where $\Phi(t_i, t_k)$ is the state transition matrix, where

$$\Phi(t_0, t_0) = I = \Phi(t_k, t_k).$$

The state transition matrix satisfies the following differential equation (11):

$$\dot{\Phi}(t, t_k) = A(t)\Phi(t, t_k), \quad \Phi(t_k, t_k) = I, \quad (11)$$

where $A(t)$ is defined in Equation 9. Using Equation 10, the system of Equations 8

can be expressed as

$$\begin{aligned} x_i &= \Phi(t_i, t_k)x_k, \\ y_i &= \tilde{H}_i x_i + \varepsilon_i, \quad i = 1, \dots, l. \end{aligned} \quad (12)$$

Now, if the first of Equations 12 is used to express each state x_i in terms of the state at some general epoch t_k , and by defining $H_i = \tilde{H}_i \Phi(t_i, t_k)$, the second of Equations 12 can be expressed (dropping the subscript i) as

$$y = Hx_k + \varepsilon. \quad (13)$$

Equation 13 represents a system of m equations in $m + n$ unknowns where $m > n$. When $\varepsilon = 0$, any n of Equation 13 that are independent can be used to determine x_k . For the general case, $\varepsilon \neq 0$ and some further criterion, such as least squares, maximum likelihood, or minimum variance must be specified to determine x_k .

Because each observation has an unknown error, there will always be m knowns (e.g., the observations y) and $m + n$ unknowns (e.g., the m observation errors ε and the n unknown components of the state vector x_k). To resolve this problem, we follow the method of least squares (12,13). In this approach, the best estimate for the unknown state vector x_k is selected as the value \hat{x}_k that minimizes the sum of the squares of the calculated values of the observation, errors, that is, if δ_k is any value of x_k , then $\varepsilon' = y - H\delta_k$ will be the m calculated values of the observation residuals corresponding to the value δ_k . Then, the best estimate of x_k will be the value that minimizes the performance index $J(\delta_k)$, where

$$J(\delta_k) = 1/2(\varepsilon'^T \varepsilon') = 1/2(y - H\delta_k)^T (y - H\delta_k). \quad (14)$$

For a minimum of this quantity, it is necessary that (7)

$$H^T H \hat{x}_k = H^T y. \quad (15)$$

Equation 15 are referred to as normal equations that represent a system of n linear algebraic equations. If the $n \times n$ matrix $(H^T H)$ has an inverse, the solution can be expressed as

$$\hat{x}_k = (H^T H)^{-1} H^T y. \quad (16)$$

In the algorithms before, it is assumed that each observation is of equal importance. If some observations are more accurate than others, the more accurate observations should be assigned a higher weight in processing the data. Typically, we use W as a diagonal $m \times m$ matrix, where a zero diagonal element implies that the corresponding observation is being neglected. This weighting matrix may also contain off-diagonal elements when the observations are correlated in some way. With this refinement, Equation 16 becomes

$$\hat{x}_k = (H^T W H)^{-1} H^T W y. \quad (17)$$

If there is an a priori estimate of the initial state, referred to as \bar{x}_0 , and, an initial estimate of the error \bar{P}_0 , then the estimate of the state deviation at t_0 can be

obtained as follows:

$$\hat{\mathbf{x}}_k = (H^T W H + \bar{P}_0)^{-1} (H^T W \mathbf{y} + \bar{P}_0^{-1} \bar{\mathbf{x}}_0). \quad (18)$$

Finally, Swerling (14) used a well-known identity to invert the final matrix on the right of the equality in Equation 18 algebraically to obtain this following sequential form of the estimation algorithm, where at time t_k ,

$$\begin{aligned} K_k &= \bar{P}_k H_k (H_k \bar{P}_k H_k^T + W_k^{-1})^{-1}, \\ \hat{\mathbf{x}}_k &= \bar{\mathbf{x}}_k + K_k (\mathbf{y}_k - H_k \bar{\mathbf{x}}_k), \\ P_k &= (I - K_k H_k) \bar{P}_k. \end{aligned} \quad (19)$$

The estimate at t_k is propagated forward to t_{k+1} by using the expressions

$$\begin{aligned} \bar{\mathbf{x}}_{k+1} &= \Phi(t_k, t_{k+1}) \hat{\mathbf{x}}_k, \\ \bar{P}_{k+1} &= \Phi(t_{k+1}, t_k) P_k \Phi^{-1}(t_{k+1}, t_k) + Q_k. \end{aligned} \quad (20)$$

For application to linear dynamic systems, the algorithm given by Equations 19 and 20 was placed on a sound statistical basis by Kalman (15) and the algorithm was extended to estimating nonlinear dynamic systems by Kalman and Bucy (16) and Bucy (17). In applying the sequential algorithm to nonlinear systems, the truncated Taylor series expansion about a reference solution, generated from complete nonlinear system dynamics, is used to obtain a linear approximation of the deviation from the reference.

Applications

During the past several decades, considerable progress has been made in the area of precision orbit determination for geodetic satellites such as LAGEOS, STARLETTE, TOPEX/POSEIDON, and ERS-1. These satellites have been used to support studies of the gravity field of Earth, its rotation and shape, ocean circulation, and tides, and the accurate determination of the location and motion of the tracking stations can be used to establish a geocentric reference frame and monitor its temporal variation (18,19). The tracking and orbit determination requirements of these missions vary considerably, but orbital accuracy is a fundamental requirement for the success of each.

Advancements in determining precise orbits received considerable stimulus from initiation of the TOPEX/POSEIDON (T/P) ocean altimeter mission (20). A description of the factors involved in this activity can be used to illuminate the applications of the POD algorithm described in the previous discussion. Accurate determination of the satellite position with respect to the Earth's mass center allows using the satellite altimeter data to monitor sea surface topography. This topography can then be combined with knowledge of the marine geoid to study the major geostrophic currents and to monitor the rise in mean sea level, both of which are important in understanding global climate change. The ability to determine the radial component of the satellite orbit at sufficient accuracy to

exploit fully the centimeter level precision of the radar altimeters requires both high-fidelity force models to describe the satellite motion and accurate tracking data that are well distributed temporally and geographically. The techniques described in the previous discussion can be used to combine these two inputs to achieve the requisite orbital accuracy.

Tracking Systems. A fundamental requirement for determining orbits that have the requisite accuracy for geodetic satellites is accurate, globally distributed tracking. In the T/P mission, three different types of tracking are available. These systems are the satellite laser ranging SLR (18), the DORIS Doppler range-rate system (21), and the GPS phase measurement system (22). Although highly accurate, the quantity and distribution of the SLR tracking data are influenced by weather, operator scheduling, and sparseness of ranging stations. The all-weather coverage provided by the DORIS system provides additional temporal and geographical coverage to complement the absolute accuracy of the SLR tracking. The combination of the SLR and DORIS tracking data sets for T/P provides nearly continuous geographical and temporal coverage from highly precise tracking systems. The GPS data collected by the GPS receiver carried onboard T/P provided highly accurate data that had very dense spatial and temporal coverage when the GPS selective availability (SA) mode was not implemented (23). This data set has provided a valuable source for model improvement and independent orbit accuracy evaluation for T/P, and the use of GPS receivers onboard satellites is expected to become widespread in the future.

The Satellite Laser Ranging System. The SLR system, which has been one of the primary geodetic tracking systems for nearly two decades, serves as the baseline tracking system for the T/P mission. The SLR measurement is the time for an optical pulse to travel from the tracking system transmitter to a reflector on the satellite and back to the tracking system. This measurement represent the state of the art in satellite tracking accuracy; its accuracy is better than one centimeter for the best instruments. In addition, the optical wavelengths are not influenced by ionospheric refraction, and the effect of water vapor is much smaller than for radiometric tracking systems. Because SLR observations provide a precise, direct measurement of the absolute range from the tracking station to a satellite, they provide good resolution of all three components of the spacecraft position with respect to the tracking network.

The DORIS Tracking System. DORIS is a one-way, ascending Doppler system that uses a set of ground beacons that broadcast continuously and omnidirectionally on two frequencies of 2036.25 and 401.25 MHz. A receiver onboard the satellite receives this signal and measures the Doppler shift, from which the average range rate of the satellite with respect to the beacon can be inferred. Average range rate is here defined as the range change across a finite count interval, usually 7 to 10 seconds for DORIS. The use of dual frequencies allows removing the ionospheric refraction. Using more than 50 beacons operating all over the world, the DORIS data provide the frequent, accurate observations of satellite motion necessary to help mitigate the effects of radiative pressure and drag forces. Thus, the DORIS system is an excellent complement to the highly precise SLR system and plays an essential role in meeting the rigorous tracking demands of the T/P mission.

The GPS Tracking System. GPS data are provided by the 24-satellite constellation of Global Positioning Satellites. This system was deployed to satisfy the real-time navigational requirements of the Department of Defense, but a number of uses of the data have been developed for highly accurate position determinations. By collecting the data using receivers located on the ground, position determination at the 1-cm level on the Earth's surface has been demonstrated. The combination of data from these accurate surface locations and the GPS measurements collected onboard the T/P satellite provides a unique data set for orbit determination or model improvement. The GPS measurement used for precise POD is the measured phase of the carrier signal, which is equivalent to a biased range measurement accurate to approximately a centimeter. Using as many as 8 to 10 GPS satellites and multiple ground stations visible to T/P at any time, a large number of observations is possible. The range measurements are biased and the clocks on the various GPS satellites and on T/P are not synchronized, but the data density provides adequate information to remove these error sources and still provide an accurate orbit for the lower satellite. The decision of the Department of Defense to allow use of GPS data at full precision provides a powerful data set for current and future geodetic missions.

Terrestrial Reference System. An important consideration in the success of the POD function is the ability to define and maintain an accurate terrestrial reference frame because this forms the reference for the location of the tracking systems described before. The international network of satellite laser ranging systems has provided the foundation for satellite geodetic advances during the past two decades. The addition of the densification from GPS and DORIS tracking systems provides input for determining and maintaining the International Terrestrial Reference Systems (ITRS). These tracking data are crucial to the current reference frame definition and provide the precise height control required to ensure the utility of the T/P results for long-term studies of ocean surface change. In the current determination of the ITRS, the accuracy of the positions of the better tracking sites is estimated at the centimeter level, both relative to the stations in the network and in an absolute sense with respect to Earth's center of mass.

Force Model Improvement. To achieve the current accuracy in POD, an intensive effort was required to improve the satellite force models, with particular emphasis on the geopotential and surface force models. To reduce the effects of atmospheric drag and the influence of errors in the ocean tide model, the TOPEX/Poseidon satellite was placed into a circular orbit at an altitude of 1336 km and inclination of 66° . For satellites at lower altitudes, atmospheric drag is an important source of orbit error and is currently limited by knowledge of atmospheric density variations, particularly when solar activity is at its maximum. For satellites where the atmospheric drag is not a problem, radiative pressure from sunlight emitted by the Sun or heat reradiated by the spacecraft become important forces.

Gravity Model Improvement. Error analysis of the best general gravity models available when the T/P mission was initiated, for example, the GEM-10B model (24), predicted radial orbital errors that approached 100 cm for the T/P orbit. Based on the observation that the gravitational model error was the primary error source, an intense effort aimed at improving the gravitational model was initiated by the TOPEX project. This effort spanned almost a decade and led

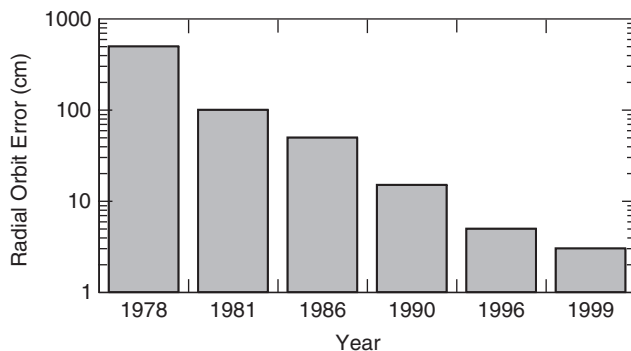


Figure 3. Progress in the accuracy of the radial orbital component since 1978 for altimeter satellites such as TOPEX/POSEIDON.

to a dramatic improvement in Earth's gravitational model, especially in the long-wavelength components that are most important to satellites. Further improvement was obtained when the GPS tracking data from the T/P satellite receiver, along with additional SLR and DORIS tracking data, were used to produce the JGM-3 gravitational model (25). Figure 3 summarizes the improvement in gravitational models for T/P from GEM-10B to JGM-3, the model currently used for T/P orbit production. For an altimeter satellite, the radial or height accuracy is the most critical to using the altimeter data.

Nongravitational Forces. Because of the significant improvement achieved in the gravitational model, nongravitational force models have become a comparable error source. The dominant effects are those due to drag and to solar, terrestrial, and thermal radiative pressure (Fig. 4). To meet current POD requirements, models are required to account for the satellite's complex geometry and attitude variations and its thermal and radiative surface properties. Typically, a relatively simple and computationally efficient model suitable for precise orbital computations is used to represent the spacecraft, although some applications require more detailed models that account for the forces acting on

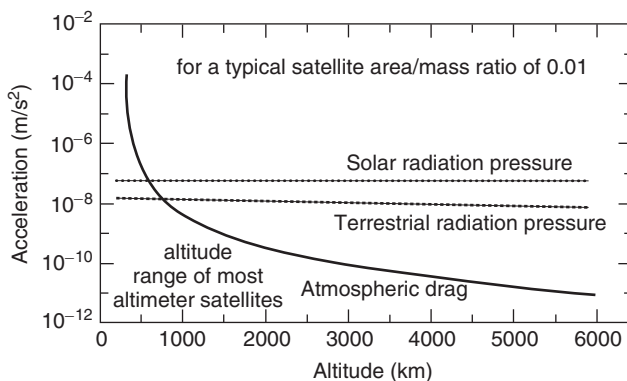


Figure 4. Relative size of nongravitational forces acting on a near-Earth satellite that has a typical area to mass ratio.

the individual surfaces of the spacecraft (26). The attitude of the spacecraft is modeled to account for the variations in area that result as the spacecraft attempts to maintain a specific orientation in space. The spacecraft attitude is especially critical for correcting the tracking measurements, such as SLR, DORIS or GPS, to the satellite's center of mass. Even though complex surface force models are employed, considerable error remains that would severely limit orbital accuracy, if not accommodated. One approach consists of estimating various empirical accelerations to absorb the errors in the surface force model that still remain (27). This allows further reduction in orbital errors due to surface force modeling deficiencies. As an alternative to modeling surface force effects, the application of three-axis accelerometers to measure surface force effects is being used on missions such as the CHAMP and GRACE gravitational mapping missions. In the future, we might expect surface-force compensation systems (so-called "drag-free" satellites) to be employed, so that the satellite orbits are entirely unaffected by nongravitational forces.

Summary

The techniques and tracking systems for computing accurate orbits for near-Earth geodetic and oceanographic satellites have undergone dramatic improvement during the last two decades. This improvement is exemplified by the orbital improvement for the TOPEX/POSEIDON orbit and for a number of other geodetic satellites. The orbital accuracies achievable when the T/P POD effort began were characterized by radial orbital errors at the meter level, but current accuracies are approaching the centimeter level. This improvement, which represents an approximately two order of magnitude increase in accuracy, has been achieved by a dramatic improvement in the accuracy and distribution of the tracking systems and the terrestrial reference frame, the use of the improved data in a sustained satellite force model development, and an extended effort to improve POD software systems. Because orbital accuracies for most geodetic satellites are routinely at the few cm level, the goal of achieving 1-centimeter accuracies for future satellite altimeter missions appears within reach.

BIBLIOGRAPHY

1. Shampine, L., and M. Gordon. *Computer Solution of Ordinary Differential Equations, The Initial Value Problem*. W.H. Freeman, San Francisco, 1975.
2. Danby, J.M.A. *Fundamentals of Celestial Mechanics*. Macmillan, New York, 1962.
3. Geyling, F.T., and H.R. Westerman. *Introduction to Orbital Mechanics*. Addison-Wesley, Reading, MA, 1971.
4. Kaula, W. *Theory of Satellite Geodesy*. Blaisdell Press, Waltham, MA, 1966.
5. King-Hele, D.G. *A Tapestry of Orbits*. Cambridge University Press, Cambridge, 1992.
6. Milani, A., A.M. Nobili, and P. Farinella. *Non-Gravitational Perturbations and Satellite Geodesy*. Adam Hilger, Bristol, 1987.
7. Tapley, B.D. Statistical orbit determination theory. In *Recent Advances in Dynamical Astronomy*, B.D. Tapley and V. Szebehely (eds), D. Reidel, Dordrecht, 1973, pp. 396–425.

8. Bierman, G.J. *Factorization Methods for Discrete Sequential Estimation*. Academic Press, New York, 1977.
9. Tapley, B.D. Fundamentals of orbit determination. In *Theory of Satellite Geodesy and Gravity Field Determination*, Vol. 25. Springer-Verlag, New York, 1989, pp. 235–260.
10. Tapley, B.D., B.E. Schutz, and G.H. Born. *Fundamentals of Orbit Determination*. Academic Press, New York, to appear 2002.
11. Coddington, E.A., and N. Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955.
12. Sorenson, H.W. Least squares estimation: From Gauss to Kalman. *IEEE Spectrum* 7: 63–68, (July 1970).
13. Lawson, C.L., and R.J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ, 1974.
14. Swerling, P. First order error propagation in a stagewise differential smoothing procedure for satellite observations. *J. Astronaut. Sci.* 6: 46–52 (1959).
15. Kalman, R.E. A new approach to linear filtering and prediction problems. *IEEE Trans. Autom. Control* 82: 34–45 (1960).
16. Kalman, R.E., and R.S. Bucy. New results in linear filtering and prediction theory. *ASME J. Basic Eng. Series D*, 83: 95–108 (1961).
17. Bucy, R.S. Nonlinear filtering theory. *IEEE Trans. Autom. Control* AC-10: 198–206 (1965).
18. Tapley, B.D., B.E. Schutz, R.J. Eanes, J.C. Ries, and M.M. Watkins. Lageos laser ranging contributions to geodynamics, geodesy, and orbital dynamics. In *Contributions of Space Geodesy to Geodynamics: Earth Dynamics, Geodynamics Series*, Vol. 24, D.E. Smith and D.L. Turcotte (eds). American Geophysical Union, Washington, DC, 1993, pp. 147–173.
19. Smith, D.E., et al. Tectonic motion and deformation from satellite laser ranging to LAGEOS. *J. Geophys. Res.* 95 (B13): 22013–22041 (1990).
20. Fu, L.L., E.J. Christensen, M. Lefebvre, and Y. Menard. TOPEX/POSEIDON mission overview. *J. Geophys. Res.* 99 (C12): 24369–24382 (1994).
21. Nöuel, F., J. Bardina, C. Jayles, Y. Labruno, and B. Troung. DORIS: A precise satellite positioning Doppler system, *Astrodynamics* 1987. *Adv. Astron. Sci.*, J.K. Solder et al. (eds) 65: 311–320 (1988).
22. Yunck, T.P., S.C. Wu, J.T. Wu, and C.L. Thornton. Precise tracking of remote sensing satellites with the Global Positioning System. *IEEE Trans. Geoscience Remote Sensing* 28 (1): 108–116 (1990).
23. Melbourne, W.G., B.D. Tapley, and T.P. Yunck. The GPS flight experiment on TOPEX/POSEIDON. *Geophys. Res. Lett.* 21: 2171–2174 (1994).
24. Lerch, F.J., C.A. Wagner, S.M. Klosko, and B.H. Putney. Goddard Earth models for oceanographic applications (GEM 10B and 10C). *Mar. Geodesy*. 5 (2): 2–43 (1981).
25. Tapley, B.D., et al. The JGM-3 geopotential model. *J. Geophys. Res.* 101 (B12): 28029–28049 (1996).
26. Antresian, P.G., and G.W. Rosborough. Prediction of radiant energy forces on the TOPEX/POSEIDON spacecraft. *J. Spacecraft Rockets* 29 (1): 81–90 (1992).
27. Tapley, B.D., et al. Precision orbit determination for TOPEX/POSEIDON. *J. Geophys. Res.* 99 (C12): 24383–24404 (1994).

BYRON D. TAPLEY
JOHN C. RIES
Center for Space Research
The University of Texas at Austin
Austin, Texas

R

ROCKET PROPULSION THEORY

Rockets

Definition. A rocket is defined as an “engine or motor that develops thrust by ejecting a stream of matter rearward, or the missile or vehicle powered by such an engine”. Since the reaction principle involved assumes a self-contained source of energy, a rocket can operate in any medium including space outside the earth’s atmosphere, where there is no oxygen to support combustion (1).

History. The Chinese are generally given credit for inventing the rocket because they appear to be the first to have employed black powder or solid rockets as weapons of war, somewhere between 1150 and 1350 A.D. (2). The Chinese attached a small rocket to the shaft of an arrow to extend its range (Fig. 1). The early history of rocket development is linked to their use as weapons; references to rockets as weapons appear from the fourteenth to eighteenth centuries (3). The technology spread to Europe and came to the attention of William Congreve of the Royal Laboratory at Woolrich, England. His stick-stabilized rocket designs were adopted by the British Navy’s arsenal and were employed in the attack on Fort McHenry in Baltimore, Maryland, that gave rise to the “rocket’s red glare” phrase in the “Star Spangled Banner.”

Modern treatments of rockets and spaceflight focus on the contributions of four men, Konstantin Eduardovich Tsiolkovsky (1857–1935), Dr. Robert Goddard (1882–1945), Dr. Hermann Oberth (1894–1989) and Dr. Wernher Von Braun (1921–1977).

The application of rocket technology to concepts of spaceflight originated with Konstantin Eduardovich Tsiolkovsky. It first appeared in 1903 in his treatise *The Investigation of Outer Space by Means of Reaction Apparatus*. A theoretician, his proficiency in mathematics and science enabled him to foresee and address such issues as escape velocities from the earth’s gravitational field,



Figure 1. Chinese fire arrow (courtesy U.S. Space and Rocket Center, Frederick I. Ordway III Collection).

gyroscopic stabilization, and the so-called “rocket equation” that establishes the relationship of the velocity increment added to a vehicle in terms of the exhaust velocity of the rocket device and the initial and final masses of the vehicle (4).

Dr. Robert Goddard was an American experimentalist who began his experiments in rocketry as a doctorate student at Clark University in Worcester, Massachusetts. His work was not widely accepted during his time. His report *A Method of Reaching Extreme Altitudes*, which offered scientifically sound concepts such as travel to the Moon, was criticized by the general public and in the press. Dr. Goddard built and successfully launched the first liquid propellant rocket on 16 March 1926. (Fig. 2). He went on to conduct additional experiments in Roswell, New Mexico, under the sponsorship of Daniel Guggenheim. He introduced the practical application of such concepts as gyroscopic stabilization of the rocket vehicle and movable deflector vanes in the rocket exhaust for directional control. He held 214 patents in rocketry (5).

Dr. Oberth was a physicist who wrote *The Rocket into Interplanetary Space* in 1923 to espouse his theories of space travel. Among his theories was the concept of staging to achieve higher velocities (6). His writings inspired Wernher Von Braun who later assisted Oberth in liquid rocket experiments. Von Braun eventually applied his knowledge to constructing liquid-fueled rocket-powered weapons during World War II. The launch of an A-4 rocket to an altitude of 50 miles (the altitude at which space is considered to begin) on 3 October 1942 might well be considered the beginning of the space age. Following the war, Von Braun and his Peenemunde team was reconstituted in the United States under the U.S. Army missile program, where its focus was again liquid-fueled rocketry. As Dr. Von Braun was developing the next generation ballistic missile, the Redstone, the United States was investing in the development of the Vanguard space launcher as a civilian launch vehicle. It was intended to inaugurate the U.S. exploration of space (7). The government intentionally avoided the use of Von Braun’s missiles as launch vehicles to emphasize the peaceful intent of space exploration. The Vanguard experienced a spectacular launch pad failure in December 1957 that

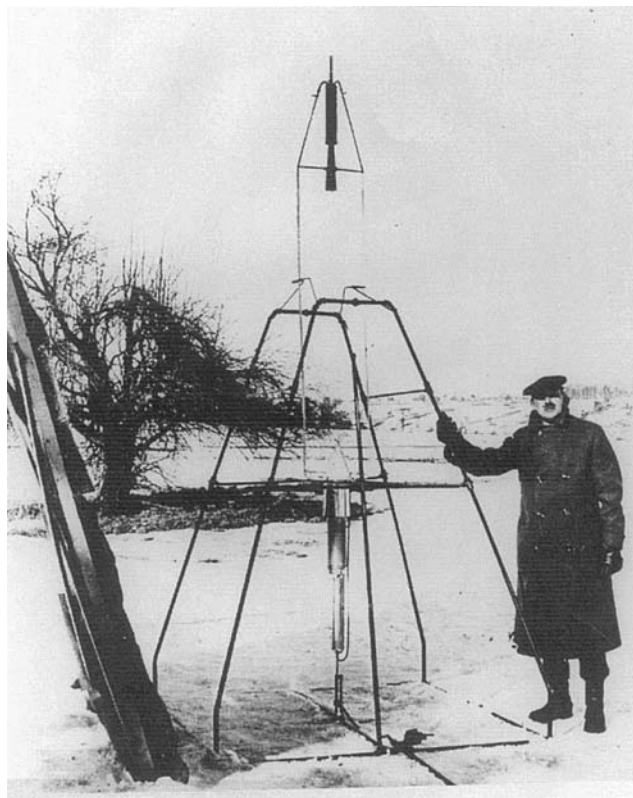


Figure 2. The Goddard rocket (courtesy U.S. Space and Rocket Center).

forced the government to turn to Dr. Von Braun in the aftermath of the successful U.S.S.R. “Sputnik” launch. His successful January 1958 launch of the Explorer 1 atop a Jupiter C launch vehicle, based on missile technology, brought Dr. Von Braun into the nation’s spotlight and resulted in his becoming the leading U.S. figure in space launch technology and space exploration. The missile technology base, thus, became the foundation for the development of space launch vehicles. As the missions became more ambitious, the need emerged for higher energy reactants than the liquid oxygen/RP-1 (kerosene) typically used, and the cryogenic system of liquid oxygen and liquid hydrogen became the standard for the civilian space launch capability. The Saturn V was the first launch vehicle developed that was not based on a vehicle or rocket engine developed for weaponry. It used liquid oxygen and liquid hydrogen in the second and third stages. Presently, the Space Shuttle, the French Arienne V, and the Japanese H-II vehicles all use hydrogen and oxygen for space launch applications.

Rocket-powered manned flight began in the 1930s in Germany. The Heinkel 176 was the first aircraft solely powered by a rocket engine. It used a 1320 lbf (297 N) engine that ran on decomposed hydrogen peroxide. In 1943, the Messerschmitt 163B rocket fighter became the first “operational” rocket-powered fighter aircraft. It had dual combustors that gave a total thrust level of 4400 lbf (989 N) and operated on hydrogen peroxide oxidizer and a fuel mixture of

hydrazine hydrate, methyl alcohol, and water. One chamber could be shut down to achieve throttling down to 660 lbf (148 N). The U.S. developments in rocket-powered manned flight began with a rocket-powered flying wing, the Northrop MX 334 (8) and eventually led to the Bell X-1 in which Captain Charles Yeager broke the sound barrier in October 1947. Ever improved experimental aircraft were flown in attempts to increase the flight speed and altitude achievable. Mach 2 was exceeded in November 1953 in a Navy Douglas D-558-II and the record was increased to 2.5 only 22 days later by Captain Yeager in the Bell X-1A. (6) The North American Aviation X-15 dominated high-speed and high-altitude research in the late 1950s and 1960s. Its 60,000-lbf (13,489 N) thrust XLR-99 engine took the X-15 to new international records for speed (Mach 6.7) and altitude (354,200 feet). Many of the lessons learned from these experimental manned aircraft were subsequently applied in designing, constructing and operating the Space Shuttle orbiter.

Solid propellant rocket motors evolved in a role as strap-on boosters to liquid stages for many space launch applications. Solid propellant rocket history was closely linked to weaponry for its development. The first large high-performance motor design was for the Polaris submarine-launched ballistic missiles where the requirements for storability and logistics of shipboard operations made the solid propellant rocket very attractive. These same requirements led to its use in the Minuteman ballistic missile. These missile developments provided the base upon which the technology for very large solid propellant rocket boosters, suitable for space launch vehicles, was built (9). The industry explored yet even larger solid propellant rocket configurations and found that their efforts were constrained by the size of the mixing facilities and the ease of transport of the rockets. These problems were overcome by using a segmented construction method in which the solid propellant rocket was cast in sections or segments, cured, then assembled into the flight vehicle at the launch site. A segmented, 156-inch-diameter solid propellant motor demonstrated the feasibility and practicality of segmented, solid propellant rocket motors, a concept employed as a strap-on booster for the Titan III launch system. This led the way to using large strap-on solid propellant rockets for the recoverable solid propellant rocket motor boosters for the Space Shuttle. A parallel development of great significance was the movable nozzle that facilitated thrust vector control for solid propellant rockets. Previously, the thrust vector for large solid propellant motors was controlled by injecting liquids into the nozzle sidewall, generating side forces by the shock caused by the interaction of the liquid jet with the supersonic nozzle flow.

Governing Laws

The operation of rocket engines and motors and the vehicles that they propel are primarily governed by Newton's laws of motion.

Newton's first law, often called the law of inertia, states that there is no change in the motion of a body unless a resultant force acts on it. A number of forces act on a launch vehicle throughout its flight. The gravitational force (weight of the vehicle), lift, drag, and the thrust of the rocket engine all act on the vehicle to cause the resultant motion. The net amount of the resultant force and its direction determine the acceleration on the vehicle and the path of the flight trajectory, in accordance with Newton's second law.

The sum of all pressures on the surfaces perpendicular to the flow axis of the device reduces to a resultant force due to the pressure differential between the pressure at the nozzle exit plane and the ambient pressure that acts on the exit area of the nozzle:

$$\text{Net pressure force} = (p_{\text{exit}} - p_{\text{ambient}})A_{\text{exit}} \quad (5)$$

The sum of the forces that act on a rocket is equal to the change of momentum in accordance with Newton's second law. By combining Equations 4 and 5 and rearranging the terms, we develop the following expression for the thrust of a rocket:

$$\text{Thrust} = \dot{m}V_{\text{exit}} + (p_{\text{exit}} - p_{\text{amb}})A_{\text{exit}}. \quad (6)$$

When p_{exit} equals p_{amb} , expansion is optimum and performance best. When the nozzle exit pressure is less than ambient, the nozzle is said to be over-expanded. If exit pressure is greater than ambient, the nozzle is said to be under-expanded. Because the rocket, generally, flies through the atmosphere, it experiences variations in the ambient pressure, so it operates at optimum expansion at only one altitude. The choice of the rocket exit area ratio then becomes the result of trading off a number of design and flight considerations.

Newton's laws are applied in analyzing the acceleration of a vehicle propelled by a rocket as well. Examining vehicle flight in a vacuum free of gravitational forces, the rocket produces an unbalanced force and a resultant acceleration in accordance with Newton's second law. Here, it is written in a way slightly different from that in Equation 2. The thrust is the net accelerating force that is equal to the instantaneous mass of the vehicle, and its instantaneous acceleration is written as the time rate of change of the velocity (10):

$$F = M \frac{dV}{dt}. \quad (7)$$

The thrust of the rocket F can also be expressed in terms of the mass flow rate \dot{m} from the rocket and its effective exhaust velocity V_e assuming that nozzle exit pressure equals the ambient from Equation 6:

$$F = \dot{m}V_e, \quad (8)$$

where

$$\dot{m} = - \frac{dM}{dt}. \quad (9)$$

The negative sign indicates that the mass of the vehicle is decreasing as the propellant exits the engine.

Substituting for thrust F from Equation 8 and mass flow rate \dot{m} from Equation 9 and rearranging,

$$dV = - V_e \frac{dM}{M}. \quad (10)$$

Integrating, we obtain the result that is commonly called the “rocket equation” that is used to calculate the velocity increment added to a stage in terms of its effective exhaust velocity and its initial and final masses:

$$\Delta V_{\text{ideal}} = V_{\text{final}} - V_{\text{initial}} = V_e \ln \frac{M_{\text{initial}}}{M_{\text{final}}}. \quad (11)$$

This solution assumed flight in a vacuum free of any gravitational field; thus, the value calculated is an ideal velocity increment. Introducing the effects of atmospheric drag and gravity result in reducing the velocity increment achieved. The gravitational field has a component of force acting along the flight path of the vehicle ($g \cos \theta$). The net loss due to the gravity field is computed by integrating this component along the path during the flight in the form of the equation,

$$\Delta V = V_e \ln \frac{M_{\text{initial}}}{M_{\text{final}}} - \int_0^t g \cos \theta dt. \quad (12)$$

The drag component is introduced as a drag coefficient C_d that is applied to the incompressible flow dynamic pressure $1/2\rho V^2$ and the cross-sectional area of the vehicle and evaluated along the flight path:

$$\text{Drag} = C_d A (1/2\rho V^2). \quad (13)$$

The drag coefficient is a function of the vehicle flight Mach number and its angle of attack (Fig. 4) (11).

Finally, the velocity increment added to a single stage during flight can be determined from the equation,

$$\Delta V = V_e \ln \frac{M_{\text{initial}}}{M_{\text{final}}} - \int_0^t g \cos \theta - \text{Drag}. \quad (14)$$

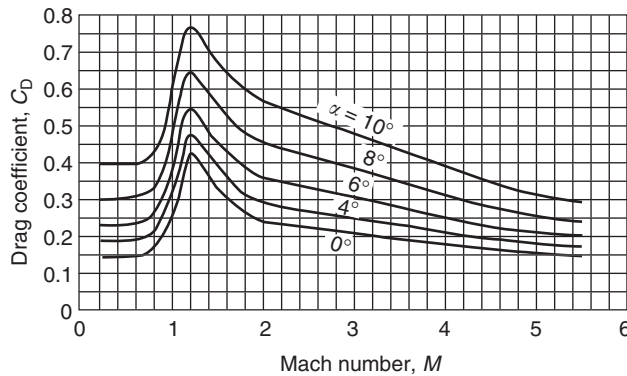


Figure 4. Drag coefficient vs. Mach number as a function of angle of attack (Reference 11, Fig. 5-3a, p. 126).

The initial mass of a stage includes the inert mass of the vehicle (structural dry mass plus engines), the mass of propellants and pressurant gases, and the payload mass:

$$M_{\text{initial}} = m_{\text{inert}} + m_{\text{propellant}} + m_{\text{payload}}. \quad (15)$$

The final mass of the stage includes the dry weight of the vehicle, any residual (unburned) propellant and pressurants, and the payload:

$$M_{\text{final}} = m_{\text{inert}} + m_{\text{residuals}} + m_{\text{payload}}. \quad (16)$$

The rocket equation can be applied to multiple-stage vehicles by solving the equation for each stage and adding the velocity increments for all stages as follows:

$$\begin{aligned} \Delta V_{\text{ideal}} = & \left[V_e \ln \frac{M_{\text{initial}}}{M_{\text{final}}} \right]_{\text{stage 1}} + \left[V_e \ln \frac{M_{\text{initial}}}{M_{\text{final}}} \right]_{\text{stage 2}} + \dots \\ & + \left[V_e \ln \frac{M_{\text{initial}}}{M_{\text{final}}} \right]_{\text{stage n}}. \end{aligned} \quad (17)$$

Care must be exercised in considering the masses of the individual stages. The payload for the first stage of a multiple-stage vehicle is the mass of the stages above it. Only in the final stage is the payload the mass that is placed in orbit.

Specific Impulse

The usual measure of performance of a rocket engine is its specific impulse. This is a measure of how much thrust (lbf or Newtons) is generated by the engine when the flow rate is 1 unit (lbm/sec or kg/s):

$$I_{\text{specific}} = I_{\text{sp}} = \frac{F}{\dot{m}}. \quad (18)$$

The units of specific impulse are seconds in SI but lbf/lbm/sec in the U.S. Customary units (USCS). “Seconds” is still used as the terminology for specific impulse in the USCS system but is not correct technically and must have the correct units for use in solving any equations. Specific impulse is related to the exit velocity of the nozzle V_e through the relationship,

$$I_{\text{specific}} = \frac{V_e}{g_c}. \quad (19)$$

Specific impulse can also be determined as a function of the properties of the working fluid and the operating conditions of the rocket nozzle. In simplified terms, we see that the specific impulse is proportional to the square root of the

temperature of the working fluid divided by its molecular weight:

$$I_{\text{specific}} = \text{const} \sqrt{\frac{T_c}{M}}. \quad (20)$$

The constant in this equation is a function of the thermodynamic properties of the combustion gases and the ratio of nozzle exit pressure to combustion chamber pressure. Therefore, the objective in any rocket engine is to achieve the highest possible temperature of the working fluid and the lowest possible molecular weight. In chemical rockets, this is typically achieved by using reactants that produce large quantities of hydrogen or steam at high temperature as their products. Beamed energy and nuclear rockets choose hydrogen for the working fluid because it is a good coolant and has the desired low molecular weight. The choice of a working fluid for electric rockets depends, in some cases, on other properties of the fluid such as ionization potential. Hydrogen is still of interest for an arcjet.

Vehicle Staging

The ideal velocity increment equation can be manipulated into the following form to examine the sensitivity of the stage performance to various aspects of rocket engine performance and stage design:

$$e^{\frac{\Delta V}{g_c I_{sp}}} = \frac{M_{\text{initial}}}{M_{\text{final}}}. \quad (21)$$

From this equation, we can see that the higher the engine specific impulse, the more closely the final mass approaches the initial mass of the stage (less propellant is consumed).

A measure of stage design efficiency is the propellant mass fraction λ defined by the equation,

$$\lambda = \frac{m_{\text{propellant}}}{m_{\text{propellant}} + m_{\text{inert}}}. \quad (22)$$

The higher the value of λ , the greater the structural efficiency of the stage. Now, we can manipulate Equations 18 and 19 into a form that presents the ratio of the initial mass of the vehicle to the payload delivered in terms of the propellant fraction, specific impulse, and ideal velocity increment:

$$\frac{M_{\text{initial}}}{M_{\text{payload}}} = \frac{\lambda e^{\frac{\Delta V}{g_c I_{sp}}}}{1 - e^{\frac{\Delta V}{g_c I_{sp}}}(1 - \lambda)}. \quad (23)$$

We can illustrate the sensitivity of stage and vehicle performance to engine specific impulse and stage structural efficiency by conducting a parametric study of two vehicles, single-stage-to-orbit (SSTO) and two-stage-to-orbit (TSTO). First, one must estimate the ideal velocity increment for the mission. A low Earth orbit typically requires about 30,000 ft/sec (9144 m/s). In the SSTO case, three specific

impulses were assumed, one representative of liquid hydrogen/liquid oxygen rocket performance (460 s), a second chosen as a conceivable increase in chemical rocket performance (500 s) and the third representative of predicted nuclear rocket performance (850 s). By assuming different values for the propellant fraction λ , one can then solve Equation 23 and arrive at the family of curves presented in Fig. 5. The lower the value of the lift-off mass to payload ratio, the better the vehicle performance. The chemical rocket options approach a limit at a value in the vicinity of 10 (10% of the vehicle liftoff mass is effective payload delivered to orbit), whereas the nuclear option approaches a limit in the vicinity of about 3 (33% of liftoff mass). The lower specific impulse vehicle, is seen as quite sensitive to the propellant mass fraction. For example, at a propellant fraction of 0.89, the 40-second difference in specific impulse between 460 s and 500 s results in a twofold difference in the initial weight to payload ratio. So, if a vehicle design were to experience an increase in weight from its initial design to completion of fabrication (and they most always do), the payload capacity of the vehicle will be severely reduced and, conceivably, could become inconsequential, depending upon the magnitude of the growth in weight. For the higher specific impulse cases, the SSTO vehicle is seen as less sensitive to any decrease in propellant mass fraction, and the nuclear rocket is virtually insensitive to propellant mass fraction across the range studied. This illustrates that the higher the delivered specific impulse, the less sensitive the stage design is to its structural efficiency.

Another way in which the initial mass/payload ratio can be made less sensitive to the propellant mass fraction is through the concept of staging in which two (or more) rocket stages can be coupled, either in parallel or in tandem. As one stage is jettisoned after exhausting its propellant load, the subsequent stage ignites and continues on to complete the mission. Figure 6 presents the results of a parametric study to illustrate the effects of specific impulse and propellant fraction upon the lift-off mass to payload ratio for a tandem two-stage vehicle that has an ideal velocity increment of 30,000 ft/s (9144 m/s). In this case, it was assumed that the propellant mass fraction was the same for each stage, and the total propellant load was distributed, 80% in the first stage and 20% in the

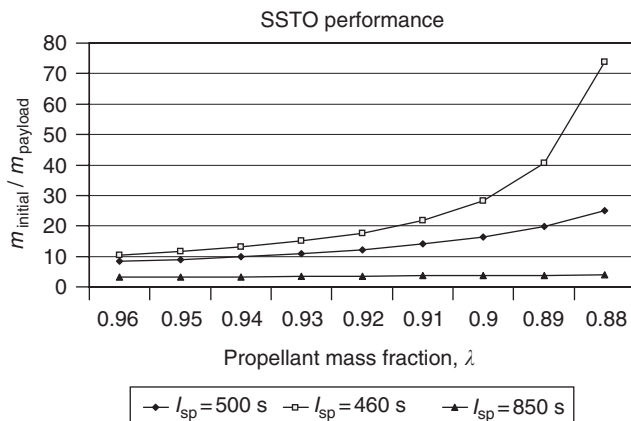


Figure 5. Payload/initial mass vs. propellant mass fraction for SSTO.

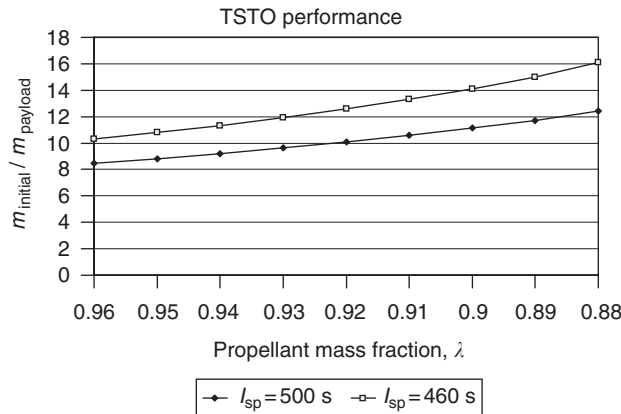


Figure 6. Payload/initial mass vs. propellant mass fraction for TSTO.

second stage. As in the SSTO case, there is a limit of initial to payload mass ratio of about 10. The maximum ratios calculated are 16 for a specific impulse of 460 s and 12 for a specific impulse of 500 s within the range of propellant mass fraction studied, as contrasted with values of 74 and 25, respectively, for the SSTO. It is easily seen that multiple stages can offer payload gains. The driver for minimizing the number of stages is cost considerations rather than performance.

Energy and Energy Conversion

Energy Conversion Mechanisms in a Rocket. The rocket engine is an energy conversion device that converts potential energy to the thermal energy of a high-temperature gas and then to the kinetic energy of a high-velocity exhaust gas. The energy sources for this conversion can be any of several types, chemical, electrical, beamed (solar or laser), or nuclear.

Chemical Rockets. In a chemical rocket, the potential energy is in the chemical bonds of the molecules that compose the fuel and the oxidizer. The class of reactants identified as “fuels” has large concentrations of hydrogen or carbon in the molecule and light metals such as aluminum, lithium, beryllium, magnesium, and boron. The class of reactants identified as oxidizers has large concentrations of oxygen, fluorine, or chlorine in the molecule (12). When the fuel reacts with an oxidizer in the combustion process, the resultant products are at a temperature significantly elevated over that at which they entered. This thermal energy is extracted for propulsion (thrust) by passing the gases through a converging-diverging (or DeLaval) nozzle that converts the thermal energy to kinetic energy. These chemical reactants are usually in either liquid or solid form.

A liquid-fueled rocket might be of either bipropellant or monopropellant type. In the former, fuel and oxidizer are introduced into the combustion device separately. They are atomized, vaporized, mixed, and combusted in the combustion chamber. The mixture ratio (proportions of oxidizer to fuel flow rate) at which the maximum combustion temperature occurs is called the stoichiometric ratio. This ratio does not yield the highest specific impulse (the ratio of thrust

produced by the engine to the rate at which propellant is being consumed), however; this occurs when using fuel-rich mixtures. The excess fuel tends to reduce the combustion temperature somewhat, but more importantly, the molecular weight of the combustion products is reduced. The net effect is to increase the specific impulse, the measure of “goodness” of rocket performance. The most common oxidizers used to date are liquid oxygen, nitric acid (HNO_3), and nitrogen tetroxide (N_2O_4). The most common fuels used to date have been liquid hydrogen, RP-1 (a kerosene-like hydrocarbon), and various amine-based fuels (hydrazine, unsymmetrical dimethyl hydrazine, monomethyl hydrazine, and mixtures of the foregoing). Liquid propellants can also be classified as cryogenics or storables. Cryogenics are liquefied gases, typically oxygen and hydrogen, and are more energetic reactants. Within the family of cryogenics, there are some propellants referred to as “space storable,” meaning that they are relatively mild cryogenics (boiling point greater than -238°F (-150 K)). These propellants can be stored in space for long periods of time and have acceptable levels of boil-off without engaging in major efforts to insulate the propellant containers (tanks) or refrigerate the propellants. Storable propellants are those that normally are liquids at standard temperature and pressure and are less energetic (13).

Monopropellants are generally introduced into a catalyst “bed” or “pack” where the propellant molecule is broken apart, liberates energy, and consequently increases in thermal energy. These reactions tend to be less energetic than bipropellant reactions and are frequently used in selected applications where specific impulse is not the most important characteristic. The most commonly used monopropellants are hydrazine (N_2H_4) and hydrogen peroxide (H_2O_2). The catalyst for hydrazine is iridium coated on an alumina substrate. The catalyst for hydrogen peroxide is silver, generally in the form of a screen.

The selection of a bipropellant combination depends on many factors, for example, the ignition characteristics of the propellant combination. Hypergolic combinations (nitrogen tetroxide and amine fuels, for example) provide a ready ignition source because the fuel and the oxidizer react upon contact. Such combinations are amenable to applications where engine restart may be required or for ignition at an altitude. Combinations such as hydrogen and oxygen require an external source of energy to provide the ignition energy for the main propellant flow. Multiple restarts are possible through appropriate design, as used on the RL-10 engine of the Centaur upper stage vehicle.

Another consideration for propellant selection is the choice of coolant for the rocket chamber. Specifically, the coolant should have a very high specific heat, low viscosity, and be thermally stable. Typically, fuels are used as engine coolants. Several designs have been tested in which oxidizers (liquid oxygen or nitrogen tetroxide) have been used as coolants (14).

The solid-propellant rocket motor is a device in which the fuel and oxidizer are premixed to form a combustible mixture which, when ignited, burns until all of the propellant is consumed. It burns at a rate that depends on the combustion pressure p and is a function of the propellant type (burn rate exponent n) according to the following relationship:

$$r = ap^n. \quad (24)$$

The burning rate r generally ranges in value from 0.3–0.5 inches per second (0.762 to 1.27 cm/s), and the burning rate exponent n ranges from 0.2–0.5 for modern propellants (15). For a detailed description of solid-propellant rockets, see the article Solid Fuel Rockets by Donald Sauvageau on page 531 of this Encyclopedia.

The hybrid is another chemical rocket embodiment. It typically consists of a solid fuel grain and a liquid (or gaseous) oxidizer. Its advertised advantages are that the fuel grain is less susceptible to safety and handling problems because the grain contains no oxidizer and, thus, cannot detonate or sustain combustion by itself. Additionally, using a liquid (or gaseous) oxidizer provides a very simple throttling scheme through oxidizer flow regulation and also provides for thrust termination by simply stopping the oxidizer flow. The fuel grain is typically designed to have several flow passages or ports through it to maximize the fuel surface exposed to the oxidizer, and, consequently, the burning surface and the attendant flow of combustion products. The hybrid has also achieved some attention because it permits eliminating the ammonium perchlorate oxidizer typical of most solid-fueled rockets and the attendant HCl in the exhaust products. The HCl presents a small but persistent environmental concern (16). The burn rate of a hybrid motor is similar in form to that of the solid-propellant rocket:

$$r = aG_o^n. \quad (25)$$

In this case, r is the burn rate, a is a constant that depends on the reactants, G_o is the oxidizer mass flow rate per port, and n is the burning rate exponent whose value is from 0.5 to 0.7. Hybrid rockets have been successfully fired in motors of up to 250,000 lbf (1.11×10^6 N) (17). The problems that exist are achieving high burning rates to minimize the size of the grain and complete consumption of the grain. Residual propellant can amount to 5 to 30% of the initial fuel load depending on the port configuration.

Nonchemical Rockets. In an electric rocket, the energy is collected from some power source (nuclear or solar) and converted to electrical energy. The energy conversion for thrusting purposes may occur by striking an arc between an anode and a cathode and passing a working fluid through the arc to heat it (arcjet). The hot gases are then passed through a nozzle to convert the thermal energy to kinetic energy. Another energy conversion technique might be to use the electrical energy to ionize an easily ionized gas (xenon, for example) and accelerate the resultant ionized particles across a potential (electrostatic thruster). Another approach may use the interaction of the current and its induced electromagnetic field that produces Lorentz forces to accelerate the gas (electromagnetic thruster) (18).

Beamed energy depends on capturing a high-energy beam from outside the vehicle and transferring its energy to a working fluid. Solar energy is one such beaming mechanism. Concentrating lenses (or mirrors) located on the vehicle capture the solar radiation and focus it on a blackbody absorber at the focal point. The absorbed energy is then used to heat a working fluid that is subsequently expelled through the nozzle at a high velocity. A laser beam might apply equally for this purpose. The working fluid also cools the absorber/thruster structure to maintain structural integrity at elevated temperatures.

Finally, a nuclear energy source may be used in place of any of the previously mentioned energy sources to heat the working fluid. Here also, the working fluid serves the dual purpose of cooling the structure and generating the thrust, as it is expelled through a nozzle.

Thermodynamics of Rockets

As previously discussed, rocket exhaust velocity is related to the specific impulse through the gravitational constant g_c as follows when the nozzle exit pressure is equal to the ambient:

$$V_e = I_{\text{specific}} g_c. \quad (26)$$

The exhaust velocity must be maximized to maximize the specific impulse. The first law of thermodynamics (conservation of energy) for bulk flow requires that the total energy entering the engine must equal the total energy leaving the engine. In equation form this is:

$$h_o = \frac{V^2}{2g_c} + h. \quad (27)$$

The average stagnation enthalpy per unit mass h_o of all of the flows entering the engine equals the kinetic energy at the nozzle exit plus the static enthalpy h of the exhaust at the exit. The static enthalpy for a chemically reacting system is given in terms of the static temperature T , entropy s and Gibbs free energy f by the equation,

$$h = Ts + f. \quad (28)$$

The Gibbs free energy is a summation of the chemical binding energy (19). Substituting and rearranging, the kinetic energy can be expressed as a function of the stagnation enthalpy Ts and the Gibbs free energy:

$$\frac{V^2}{2g_c} = h_o - Ts - f. \quad (29)$$

By differentiating the kinetic energy with respect to entropy s and Gibbs free energy f , we can find that the exhaust velocity for a given exit pressure increases as exit entropy decreases and Gibbs free energy decreases. From the second law of thermodynamics, the minimum exit entropy is equal to the inlet entropy for bulk flow. Therefore, the maximum exhaust velocity occurs for an exit entropy per unit mass that is equal to the inlet entropy per unit mass. Additionally, the minimum Gibbs free energy occurs under conditions of chemical equilibrium. Chemical equilibrium is achieved as the chemical constituents of the exhaust gases alter their relative proportions in response to the pressure and temperature changes that occur as the gases flow through the nozzle. Energy is liberated to the expansion process as a result. Therefore the maximum specific impulse is achieved for an ideal one-dimensional flow when the exhaust products

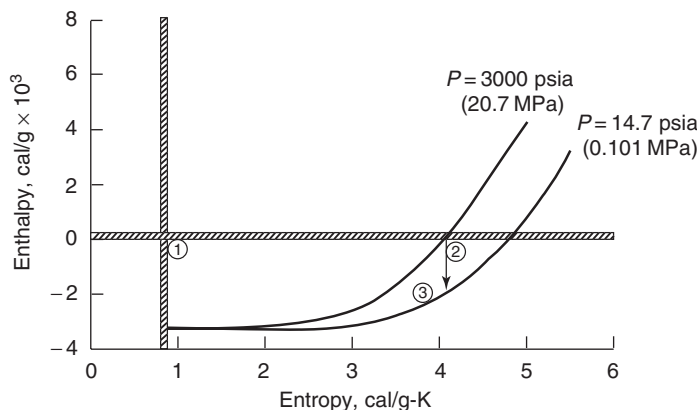


Figure 7. Enthalpy vs. entropy for flow in a rocket engine (courtesy Boeing, Rocketdyne Propulsion and Power).

are expanded to ambient pressure in chemical equilibrium and have total enthalpy and entropy equal to the inlet conditions (20).

The reality of rocket operation is that none of the processes occurs ideally. The combustion process is irreversible and has entropy increases attendant to it. Viscous boundary layer effects and nonequilibrium expansion of the combustion products also introduce irreversibilities. Figure 7 presents the enthalpy versus entropy diagram for an oxygen–hydrogen reaction in chemical equilibrium at 3000 psia (20.7 MPa). Process 1 to 2 is the irreversible combustion of oxygen–hydrogen reactants at 3000 psia (20.7 MPa). The equilibrium products are then expanded in a constant entropy (isentropic) process to local ambient pressure (0.101 MPa) from point 2 to point 3.

Nozzle Theory

The nozzle is the mechanism for accelerating the hot gases in all of the devices described, except those that use electrostatic or electromagnetic forces. Applying the first law of thermodynamics across the nozzle, the change in enthalpy equals the kinetic energy of the exhaust gases at the exit:

$$\frac{V_{\text{exit}}^2}{2g_c} = h_o + h_{\text{exit}}. \quad (30)$$

Applying the idealization of isentropic, one-dimensional flow, the exhaust velocity of the nozzle is

$$V_{\text{exit}} = \sqrt{\left(\frac{2\gamma}{\lambda - 1}\right) \left(\frac{\bar{R}}{M}\right) T_c \left[1 - \left(\frac{p_{\text{exit}}}{p_{\text{chamber}}}\right)^{\frac{\gamma-1}{\gamma}}\right]}. \quad (31)$$

The nozzle is typically identified by its geometric area ratio (exit area/throat area), but the pressure ratio across the nozzle is important in determining its

performance in accelerating the gas flow, as seen by examining the equation for exit velocity. The term γ is the ratio of specific heats of the hot gases, " \bar{R} " is the universal gas constant, and M is the molecular weight of the gases that exit the nozzle. Optimum expansion of the gases occurs when the exit pressure of the nozzle equals the local ambient pressure.

The typical shape of a DeLaval nozzle is shown in Fig. 8. In the subsonic region (flow velocities less than Mach 1), the gases are accelerated by decreasing the area of the flow passage. Continuing the decrease of the flow area increases the gas velocity until a point is reached at which the maximum mass flow rate per unit area is achieved. At this condition, the flow is at the speed of sound or sonic (Mach number equal to 1). This location is called the throat of the nozzle, and the flow is referred to as "choked." The ratio of chamber pressure to pressure at the throat (critical pressure ratio) is approximately 2:1 for this condition. From that point on, the flow passage must increase in area to permit continuing acceleration of the flow in the supersonic regime (Mach numbers greater than 1). Once the nozzle achieves the choked condition, the chamber pressure remains

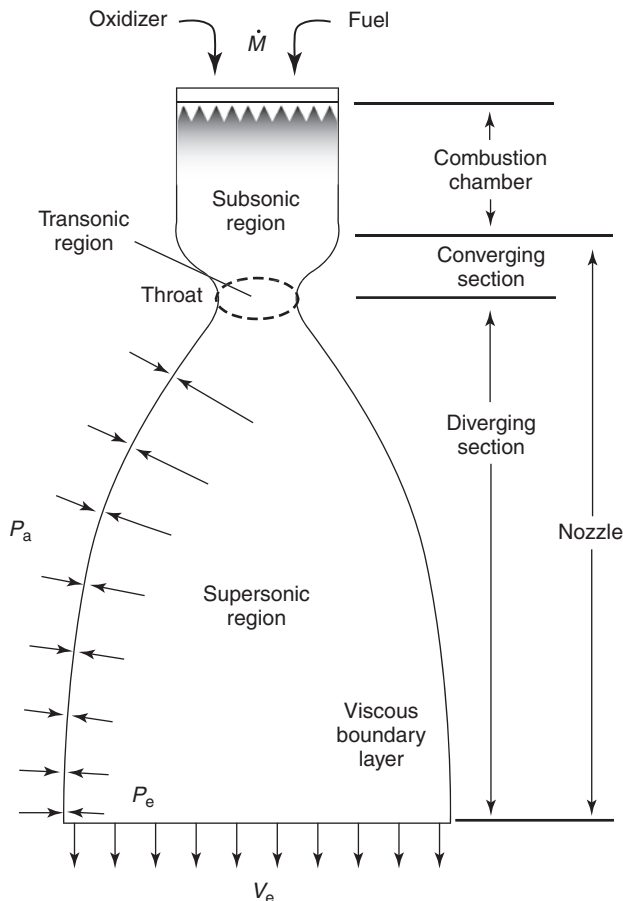


Figure 8. The DeLaval Nozzle (courtesy Boeing, Rocketdyne Propulsion and Power).

constant regardless of the back-pressure from the flight altitude. If the exit pressure exceeds the local ambient, it is underexpanded; if it is less than the local ambient, it is overexpanded. Selection of the nozzle area/pressure ratio is a compromise to provide the best performance across the vehicle's flight regime. One way to examine this design choice is through the nozzle thrust coefficient C_f . The thrust coefficient is a measure of nozzle performance and can be used to determine the thrust of a rocket engine as a function of the throat area and the chamber pressure from the equation,

$$F = C_f A_t p_c. \quad (32)$$

For a given ratio of specific heats, the optimum geometric area ratio corresponds to a given pressure ratio across the nozzle. For example, at a nozzle area ratio of 10, the design pressure ratio for a ratio of specific heats of 1.3 is 100, and the corresponding thrust coefficient is 1.6. If a rocket has a nozzle of area ratio of 10, it is operating at its optimum condition only when the pressure ratio is 100. Because a rocket flies through the atmosphere and is subject to varying ambient pressure, the delivered performance will be less than optimum at all other points in the trajectory than when the pressure ratio is 100. Therefore, the choice of design area ratio (pressure ratio) is a compromise in which the designer knowingly accepts less than optimum performance at some portions of the trajectory.

The nozzle designer must make compromises in choosing the design pressure ratio and also in the shape of the nozzle. However, the most common nozzle design practice uses the Rao optimum contour nozzle also known as the "bell" nozzle (21). This design yields a shorter design than a simple 15° half-angle conical shape and has lower divergence losses because the gases are exiting the nozzle at a divergence angle of less than 8° .

The losses in a rocket nozzle consist of the divergence loss (nonaxial velocity vector for the exiting gases), finite-rate kinetic losses, and drag losses, in accordance with the equation,

$$\eta_{\text{nozzle}} = \eta_{\text{divergence}} \eta_{\text{kinetic}} (1 - \eta_{\text{drag}}). \quad (33)$$

Inserting typical values into this equation, we can find the overall efficiency of a modern rocket nozzle design (22):

$$\eta_{\text{nozzle}} = (0.992)(0.999)(1 - 0.986) = 0.997. \quad (34)$$

There are nozzle designs that are intended to compensate for altitude as the rocket flies through the constantly varying pressure of the atmosphere. This includes mechanical means in which nozzle "skirts" are moved into place at selected times in the flight profile; each increases the area/pressure ratio to approximate optimum expansion more closely. There are also aerodynamic designs such as the "aerospike" shown in Fig. 9 that have a free jet boundary that adjusts to the local ambient pressure through a Prandtl–Meyer or corner expansion (23). Such an expansion process constrains the nozzle gases to expand to the local ambient at the outside lip of the nozzle. Consequently, the flow cannot overexpand, and the nozzle flow is always optimum until the nozzle design pressure ratio is exceeded. In practice, these nozzle designs do not operate with

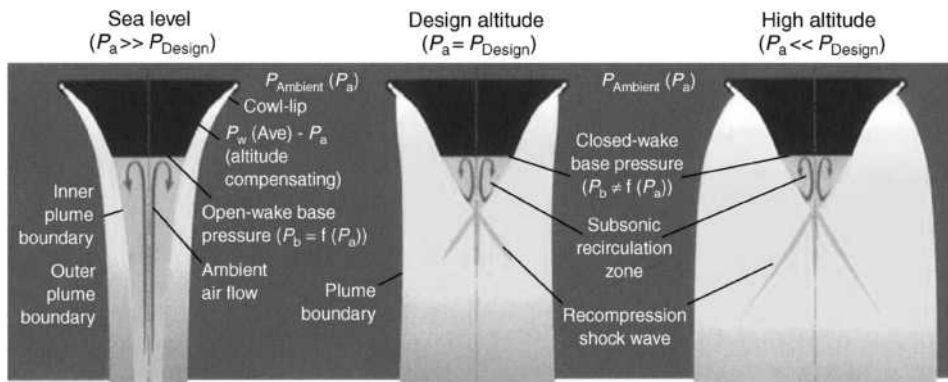


Figure 9. Aerospike nozzle and its operating modes (courtesy Boeing, Rocketdyne Propulsion and Power). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

perfect altitude compensation. A procedure has been developed to determine the degree to which the aerospike type of nozzle approaches optimum expansion. This procedure is illustrated in Figs. 10 and 11. Figure 10 presents the thrust efficiency C_T (the ratio of measured C_F to ideal C_F) of both an altitude compensating and bell nozzle versus pressure ratio. Both nozzles have the same area ratio (design pressure ratio), the same thermodynamic properties of the flowing gases, and the same maximum thrust efficiency. At a given pressure ratio, the percentage of maximum thrust efficiency C_T achieved by the altitude compensating nozzle is compared to that of the fully flowing bell nozzle. The results of this comparison are presented in Fig. 11. The 45° line across the graph represents the performance of a bell nozzle in which the flowing gases do not separate. All points above that line represent some degree of altitude compensation. The point identified by $C_{T\text{ADV}}/C_{T\text{MAX}} = 1$ and $C_{T\text{N-S}}/C_{T\text{MAX}} = 1$ represents the design

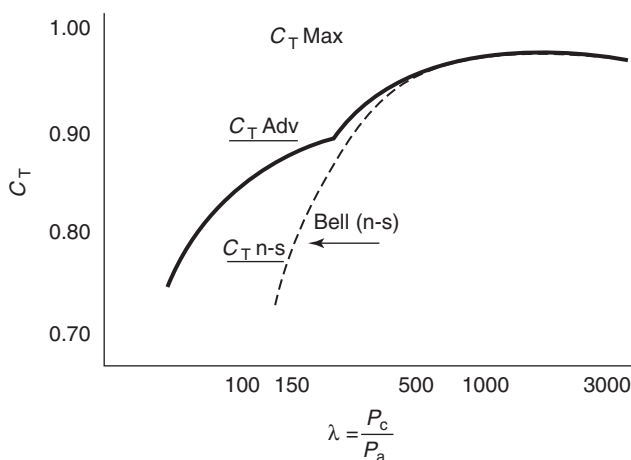


Figure 10. Nozzle altitude compensation determination.

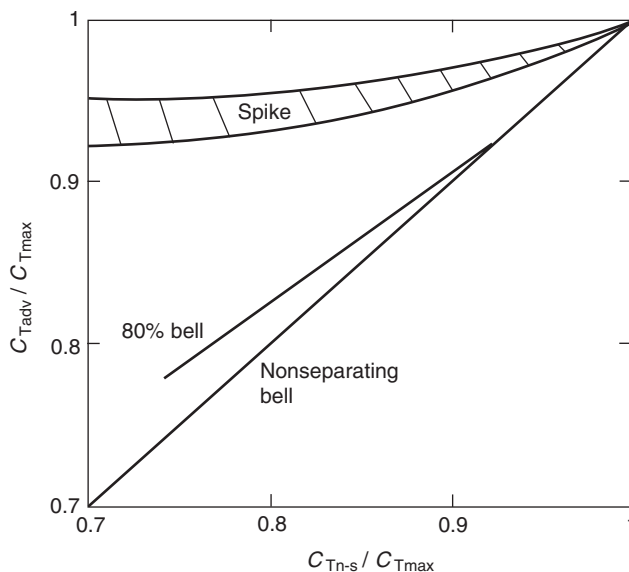


Figure 11. Nozzle altitude compensation comparison.

point for both nozzles. The line where $C_{TADV}/C_{TMAX} = 1$ represents perfect altitude compensation. The crosshatched area is representative of the altitude compensation performance of the aerospike type of nozzle. In general, the data indicate a compensation capability in excess of 50%, as determined by the ratio of the distance from the aerospike family of curves to the line where $C_{TADV}/C_{TMAX} = 1$ divided by the distance from the 45° line (nonseparating bell performance). The line identified as the “80% bell” illustrates the fact that a bell nozzle can have some separation of the flow at the low end of its operating regime, and the result is that it has some altitude compensation. The breadth of the aerospike band results from the range of nozzle lengths from 16% of an isentropic spike at the upper end to 0% at the lower end.

Additionally, the aerospike nozzle length is about 25% that of the Rao optimum nozzle length for the same thrust capability. This allows reducing in vehicle length or including more propellant in the vehicle.

The same Prandtl–Meyer expansion process occurs in the external flow aft of the flight vehicle that produces a pressure at the nozzle lip (the flow controlling pressure) that is lower than the ambient at the flight altitude. Thus, the nozzle behaves as though it were operating at an altitude higher than actual. This is most noticeable in the Mach 1 to Mach 3 range of flight operations. In that flight regime, the time spent in off-design conditions is considered negligible compared to bell nozzle performance capabilities. The added performance of optimized nozzle expansion across the full flight trajectory opens the possibility of single-stage-to orbit operations.

Rocket Engine Efficiencies. The overall efficiency of a rocket engine is generally presented as a percentage of the theoretical specific impulse. The calculated theoretical value is based on shifting equilibrium, a case in which the reaction products remain in equilibrium (their proportions change) as the

pressure and temperature decrease during the expansion process in the nozzle. This efficiency is composed of the combustion and nozzle efficiencies, generally, it is greater than 90% and can be as high as 95–96%. The nozzle efficiency depends on design, but can be expected to be close to the 98% value previously shown. The combustion efficiency of the propellants is quite sensitive to the propellant combination used and plays the major role in overall efficiency. The liquid oxygen/hydrocarbon combination of the early space launch engines typically had combustion efficiencies in the low 90s% range. The efficiencies of today's hydrogen/oxygen engines are much higher and approach 97%.

BIBLIOGRAPHY

1. *The Illustrated Columbia Encyclopedia*, Vol. 17. Columbia University Press, New York, London, 1963.
2. Needham, J. *Science and Civilization in China*. Cambridge Press, Cambridge, 1988, pp. 472–520.
3. Konstam, A. The war rocket in late Victorian military service. *Ordinance Soc. J.* 2: 51–70 (date unknown).
4. Anonymous, "Rocketry" [online], 1999, Available www.allstar.fiu.edu/aero/tsiolkovsky.htm
5. Anonymous, Goddard, Robert, [online], 1999, Available www.treasure-troves.com/bios/Goddard.html
6. Anonymous, The History of Aviation and Modern Rocketry [online], 1998. Available www.thespaceplace.com/history/rocket2.html
7. Johnson, T. Vanguard Launch Vehicle [online], Available www.76.pair.com/tjohnson/vangrd.html
8. Ehresman, C.M. Liquid rocket propulsion applied to manned aircraft in historical perspective. AIAA Paper 91-2554, Monterrey, June 1991.
9. Gavin, J.G. et al. *From Earth to Orbit, An Assessment of Transportation Options*. National Academy Press, Washington, DC, 1992.
10. Barrère, M., A. Jaumotte, B. Fraeijs De Veubeke, and J. Vandenkerckhove. *Rocket Propulsion*. Elsevier, Amsterdam, 1960, pp. 712–714.
11. Sutton, G.P. *Rocket Propulsion Elements*, 6th ed. Wiley, New York, 1992, p. 126.
12. Barrère, M., A. Jaumotte, B. Fraeijs De Veubeke, and J. Vandenkerckhove. *Rocket Propulsion*. Elsevier, Amsterdam, 1960, p. 587.
13. Webber, B., and F. Gunderloy. What's in the Tank: A Propellant primer. *Threshold 4*: (Spring 1989) Rockwell International, Canoga Park, CA, pp. 35–42.
14. Hawk, C.W., and R. Sekita, Lessons learned in the development of staged combustion liquid rocket engines. ISTS paper 94-a-10v, *19th Int. Symp. Space Technol. Sci.*, Yokohama, Japan, May 1994.
15. Sutton, G.P. *Rocket Propulsion Elements*, 6th ed. Wiley, New York, 1992, p. 375.
16. Hawk, C.W. Environmental Aspects of Rocket and Gun Propulsion. AGARD-CP-559, Alesund, February 1995.
17. McFarlane, J.S., R.J. Kniffen, and J. Lichtowich, Design and testing of AMROC's 250,000 lbf thrust hybrid motor. AIAA Paper 93-2551, *AIAA/SAE/ASME/ASEE 29th Joint Propulsion Conf. Exhibit*, Monterey, CA, June 28–30, 1993.
18. Jahn, R.G. *Physics of Electric Propulsion*. McGraw-Hill, New York, 1968.
19. Russell, L.D., and G.A. Adebisi. *Classical Thermodynamics*. Saunders, Fort Worth, TX, 1993, p. 724.

20. Evans, S.A., and B.J. Waldman. The Thermodynamic Basis of Rocket Engine Performance. Rocketdyne, North American Rockwell, July 1970.
21. Rao, G.V.R. Exhaust nozzle contour for optimum thrust. *Jet Propulsion* 28: 377–382 (June 1958).
22. O'Leary, R.A., and J.E. Beck. Nozzle Design. *Threshold* 8: (Spring 1992), Rockwell International, Rocketdyne Division, Canoga Park, CA.
23. Shapiro, A.H. *The Dynamics and Thermodynamics of Compressible Fluid Flow*, Vol. I. Ronald Press, New York, 1953, p. 463.

READING LIST

- Oberg, J.E. *Red Star in Orbit*. Random House, New York, 1981.
- Holtzmann, R.T. *Chemical Rockets and Flame and Explosives Technology*. Marcel Dekker, New York, 1969.
- Clark, J.D. *Ignition! An Informal History of Liquid Rocket Propellants*. Rutgers University Press, New Brunswick, NJ, 1972.
- Davenas, A. (ed.). *Solid Rocket Propulsion Technology*. Pergamon, Oxford, England, 1988.
- Huzel, D.K., and D.H. Huang. *Modern Engineering for Design of Liquid-Propellant Rocket Engines*. American Institute of Aeronautics and Astronautics, Washington DC, 1992.
- Clafin, S., and J. Volkmann. The Emergence of hybrid rockets. *Threshold*, 11: (Winter 1993) Rockwell International, Canoga Park, CA, pp. 33–39.
- Beck, J.E., and M.D. Horn. Altitude compensating nozzles. *Threshold* 13: (Summer 1995) Rockwell International, Canoga Park, CA, pp. 38–44.
- Klager, K. The propellant chemists' contribution to modern rocket flight—A personal memoir. IAF Paper 86-491, 37th *Int. Astronaut. Congr.*, Innsbruck, Austria, October 4–11, 1986.
- Clark, P., and K.B. Hindley. *USSR Rocket Engines*, 2nd ed., Technology Detail, York, UK, January 1992.

CLARK W. HAWK
Madison, Alabama

ROCKETS, ION PROPULSION

Introduction

Electric propulsion, the propelling of space vehicles by streams of electrically charged and electrically accelerated particles, is one of the most exciting concepts of the space adventure. It represents a fascinating area of applied science and technology, and it also provides the only feasible means of exploring the farthest reaches of our solar system by economically transporting large manned expeditions and heavy payloads to neighboring planets and probing the solar system away from the ecliptic plane.

The concept of the ion rocket engine (the most useful form of electric propulsion for space missions) has existed for some time. R.H. Goddard alluded to the possibility by an entry in his notebook dated 1906, and in 1929, H. Oberth

discussed it in his classic book *Wege zur Raumschiffahrt* (1). However, the concept had to remain of only academic interest until the advent of the space age, and, particularly, until generating large amounts of electric power in space became a physical possibility. When solar-electric power source technology was established, the latter requirements, of course, became fulfilled.

In the United States, active interest dates back to about 1954 when NASA's Dr. Ernst Stuhlinger published some comprehensive space system studies on the subject of electric propulsion (2). The first experiments on ion engines started about 1958, and an accelerated effort continued through the 1960s and 1970s. The ion engine matured during the 1980s and 1990s, and functional satellite and deep-space missions occurred in the late 1990s.

In general, an electric propulsion system has a power plant that consists of an energy source and a means of converting this energy into electrical energy and a propulsion device that converts the electrical energy into directed kinetic energy of its propellant. The two major problem areas of electric propulsion are 1) the development of sources of electrical energy that are sufficiently lightweight and 2) the transfer of this energy to the propellant to provide high exhaust velocities efficiently. The first problem, along with the general need for power in space, has given strong impetus to fundamental and applied work in developing lightweight solar panels. These areas of research and development have been vital to the future of electric propulsion and obviously have impacted many other important satellite and spacecraft applications. A solution to the second problem, an efficient propulsion device, is the subject of this article (3).

Relationship to Chemical Rockets

Before proceeding with the discussion of ion propulsion, it might be well to relate it to the more conventional types of rocket propulsion, both in terms of its basic mechanism and of the types of missions for which it is required. Thrust is equal to the rate of change of momentum imparted by the exhaust of a rocket engine. For simplicity, let it be assumed that all particles in the exhaust have the same velocity v_{ex} . Then the thrust T is simply

$$T = \dot{M}_p v_{\text{ex}}, \quad (1)$$

where \dot{M}_p is the propellant mass flow rate. Thus, if we can increase the exhaust velocity, a smaller expenditure of propellant is required for a given thrust. From a slightly different point of view, consider the equation of motion in free space for a vehicle whose mass M and velocity v are functions of time:

$$M dv = -v_{\text{ex}} dM. \quad (2)$$

Hence, for an initial mass M_i , starting from rest,

$$v = v_{\text{ex}} \ln M_i/M. \quad (3)$$

This relation tells us that the effectiveness of mass expenditure ($M_i - M$) to gain velocity is a function of exhaust velocity, that is, from a mass point of view, a

given velocity increment is achieved more efficiently from a system that has high exhaust velocity.

Now, in a chemical rocket engine, burning a fuel generates thrust; the hot gases from the combustion process are expelled through a nozzle at an exhaust velocity related to the temperature of the burning propellant. This temperature, in turn, is limited by the chemical energy content of the fuel. Thus, in a chemical system, exhaust velocity is low and consequently, from equation 1, the mass flow rate is very high. These systems, of course, have high thrust and are needed to overcome gravitational forces and to inject payloads into orbit. However, for many projected space missions starting from an Earth orbit or after Earth escape, the tremendous mass of a chemical rocket that would be necessary to deliver a heavy payload to, say Jupiter, may be completely unfeasible.

It is evident that the acceleration of ions (e.g., in ion thrusters) to any desired velocity is completely unrelated to thermal heating: the exhaust velocity can be adjusted to any desired high value, resulting in great fuel economy. We replace the high-thrust chemical systems and their huge weight of propellants by low-thrust electrostatic systems that operate for extremely long periods of time and require much lighter fuel loads.

Specific impulse is a measure of exhaust velocity. If a thrust T lasts for a time τ , the total impulse imparted to the space vehicle is $T\tau$. Then, the specific impulse, defined as impulse per unit weight of propellant exhausted, is

$$I_{sp} = T\tau/W_p = T/\dot{W}_p, \quad (4)$$

where $\dot{W}_p = M_p g$ is the propellant weight. Combining equations 1 and 4, we obtain

$$I_{sp} = v_{ex}/g \text{ seconds } (g = 9.8 \text{ m/s}^2), \quad (5)$$

which identifies specific impulse as a measure of exhaust velocity.

The best chemical systems have specific impulse values slightly more than 300 s, whereas ion rocket engines efficiently provide values of specific impulse from around 3000 s on up—a factor of 10 or more better than the most advanced chemical systems.

For any particular space mission, an optimum value of specific impulse (or exhaust velocity) will always exist that results in a minimum overall system weight (viz., useful payload, engine, and propellant), which must be boosted into the initial Earth orbit or beyond. This concept is illustrated in Fig. 1 and arises as follows. At low I_{sp} , the propellant mass is high, whereas at high I_{sp} , the weight of the electric power supply increases (power-supply weight is roughly proportional to power and, in turn, the power in the exhaust beam is proportional to the product of thrust and exhaust velocity). Thus, the optimum specific impulse involves a compromise between power-system weight and propellant weight and is larger, the more energetic the mission. From Fig. 1, we also see that the optimum specific impulse depends on the slope of the power plant weight curve, that is, the power plant weight per unit power output α_p . This parameter is commonly designated power plant “specific weight” and, in general, the lower the “specific weight” of the power plant for a given mission, the higher the optimum

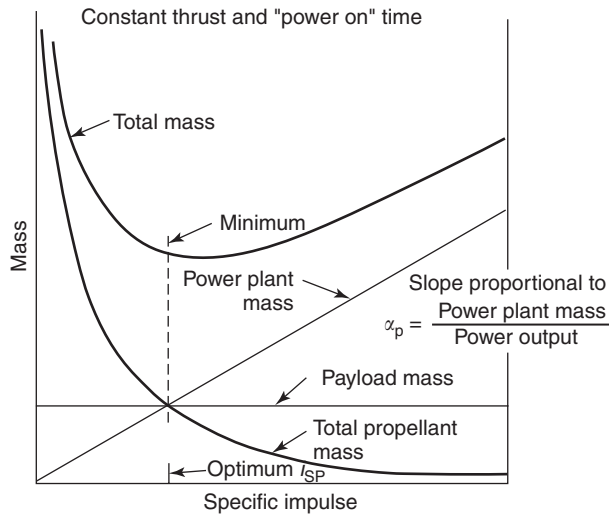


Figure 1. Mass of a complete ion propulsion system.

specific impulse. For ion propulsion systems powered by solar panels, the optimum specific impulse falls into the range of 3000–4000 s.

The low thrust achievable from ion propulsion results from the fact that such systems are power-limited. A power-to-thrust ratio of about 100 kW/lb thrust is representative of the current art. This figure will gradually decrease to around 75 kW/lb as the technology advances. Thus, for example, power at the several-hundred-watt levels limits the thrust to millipounds. However, when gravitational forces are small (such as for interplanetary spacecraft that have been launched beyond the gravitational pull of Earth) or when they are balanced by centrifugal forces (as in satellites), total impulse replaces thrust as a fundamental requirement. For example, a thrust of one pound that acts continuously for one year produces the same total impulse as a thrust of 500,000 lb that acts for one minute. Therefore, low thrusts, applied for a long period of time and

Table 1. Comparison Chart

Spacecraft description	All-chemical voyager (4), 190 days transit	SEP spacecraft, ^a 250 days transit
Injected weight (not including launch vehicle)	7800 lb	9600 lb
Power level	All chemical	23 kW
Approach velocity	4.3 km/s	1.8 km/s
Weight at approach	5295 lb	5140 lb
Weight in orbit (excluding retro inert weight)	1850 lb	3598 lb
Orbit spacecraft fraction	0.35	0.70
Lander weight	2300 lb	2300 lb
Scientific payload	470 lb	1793 lb
Percent scientific payload weight at approach	8.9%	34.9%

^aSolar electric propulsion.

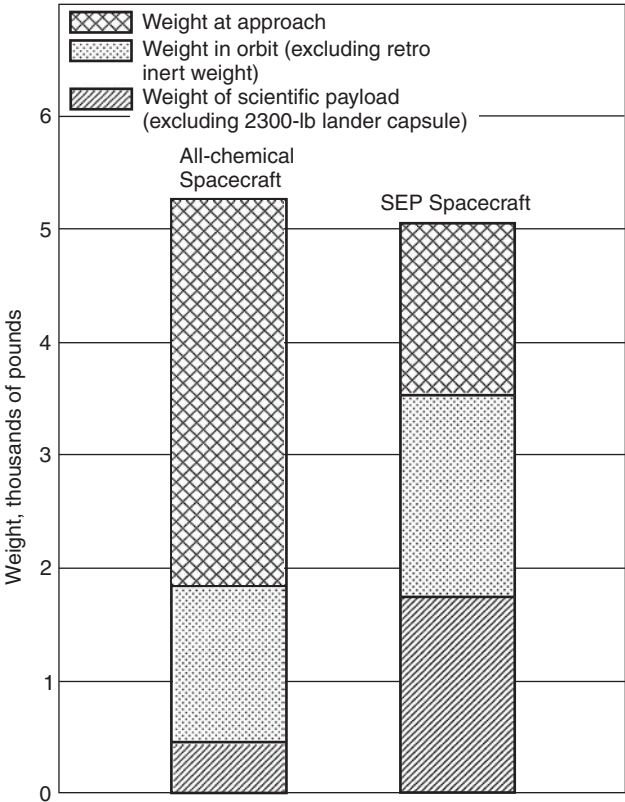


Figure 2. Performance comparison.

coupled with great fuel economy (high I_{sp}) can make possible a number of long-duration space missions that otherwise would not be feasible.

A Mars orbiter mission was selected as a basis of comparison between an all-chemical propulsion spacecraft mission and one in which chemical propulsion is augmented by an appropriate solar-powered electric propulsion system. A 2300-lb lander capsule is carried by both spacecraft and it is assumed, separates from the bus before retro into orbit. The weight breakdowns are presented in Table 1. The results (see Fig. 2) indicate that although the weight at approach is comparable for both spacecraft, the electric propulsion spacecraft places useful weight in orbit almost *twice* that of the all-chemical spacecraft. More significant is the fact that the actual scientific payload placed in Mars orbit by the electric propulsion system is nearly four times that of the all-chemical spacecraft.

Generic Types of Electric Propulsion Thrusters

We have compared ion propulsion to chemical propulsion from a performance standpoint. Now, let us briefly look at the technological differences among the various forms of electric propulsion devices.

General Principles. Over the years, there have been varying approaches to the electrical generation of thrust, characterized by their acceleration mechanism. These thrusters are classified as 1) electrothermal or thermally heated (5), 2) electromagnetic or plasma (6), and 3) electrostatic or ion (7).

Electrothermal. Electrothermal thrusters generate thrust by increasing the enthalpy of the propellant (such as hydrogen) by raising it to very high temperatures. The gas is then expanded through a nozzle where its thermal energy is converted into directed kinetic energy.

Electromagnetic. Electromagnetic thrusters generate thrust by ionizing the propellant and driving high currents through the resultant plasma. The current-carrying plasma conductors interact with the self-field of the net current in the plasma and discharge electrodes to accelerate the entire plasma. Consequently, electromagnetic thrusters are called plasma thrusters.

Electrostatic. This article deals with electrostatic or ion thrusters that generate thrust by ionizing the propellant gas and then by accelerating ions in an electrostatic field. It is more advanced than plasma propulsion in the sense that much progress has been made in solving the basic problems (such as efficiency and life) and that ion engines are presently being manufactured and are available for actual high-energy missions.

Specific Examples of Generic Types. The three basic electric-propulsion types generally described before are further sub divided into a number of specific devices. A partial listing of these devices is presented in Table 2. Figure 3 shows an example and some design specifics for each generic type of electric propulsion thruster. Now, a brief description of each.

Resistojet. In a resistojet, the propellant is heated by energy transferred from an electrically heated element (resistor). One class of resistojet, the augmented hydrazine thruster, uses electric power and/or catalyst beds to initiate and sustain decomposition of the hybrid hydrazine. The vaporized hydrazine is then passed through a heat exchanger where additional energy is added to the propellant. Resistojets achieve a specific impulse of about 300 s and

Table 2. **Electric-Propulsion Thrusters**

Electrothermal technologies
<ul style="list-style-type: none"> • Resistojet • Arcjet
Electromagnetic technologies
<ul style="list-style-type: none"> • Magnetoplasmadynamic (MPD) thruster • Pulsed plasma thruster • Pulsed inductive thruster (PIT) • Rail accelerators
Electrostatic technologies
<ul style="list-style-type: none"> • Electron-bombardment ion thrusters • Hall-effect ion thrusters • Surface-contact ion thrusters

a thrust-to-power ratio of about 575 mN/kW; this makes them well suited for low Δv applications, such as drag makeup and modest orbit raising. Approximately 200 resistojets have been flown on US satellites since 1983.

Arcjet. In the arcjet, the propellant stream is heated by an electric discharge arc struck between the thrust chamber wall (i.e., the anode) and a tungsten cathode. As in the resistojet, the hydrazine arcjet uses a catalyst bed to decompose the propellant before it enters the arc chamber. Low-power arcjets achieve an I_{sp} of about 500 s and produce a thrust of ≈ 230 mN using input power of 1.8 kW. Although a high-power (≈ 30 kW) arcjet has been demonstrated in a short-duration space test, the most prominent use of the arcjet has been for North-South stationkeeping on communications satellites. Approximately 100 arcjets have been launched on commercial satellites since 1993.

Magnetoplasmadynamic (MPD) Propulsion. The electrode and discharge configuration of a self-field MPD thruster is shown in Fig. 3. Gaseous propellant is introduced between the cathode and anode, and a voltage pulse is applied to ionize the gas and produce current flow between cathode and anode, as shown. The current that interacts with the self-fields, as shown, produces force on the conducting plasma. The nozzle-like shape of the anode also contributes some thrust as a consequence of the ohmic heating of the propellant gas by the high discharge current (10^4 to 10^5 A) and subsequent expansion of the hot gas as it flows through the nozzle.

Hall-Effect Ion Thruster. The Hall-effect thruster, also called the closed drift or stationary plasma thruster, has been under intensive development by the Soviet Union during the past four decades. The Hall thruster accelerates electrons emitted by an external hollow cathode into an annular ionization chamber, which has an anode located at its upstream end. The axial motion of the electrons across an applied radial magnetic field gives them a drift velocity in the azimuthal direction. The electrons drift in closed paths and produce ions by colliding with xenon gas atoms, hence the name closed-drift thruster. The electric field that accelerates the electrons in the upstream direction also accelerates the positive ions in the downstream direction to produce thrust. Typical Hall-effect thrusters achieve an I_{sp} of ≈ 1600 s and a thrust of ≈ 83 mN, when operating at an input power of 1.35 kW. The Soviet Union has launched approximately 100 Hall-effect thrusters on GEO satellites.

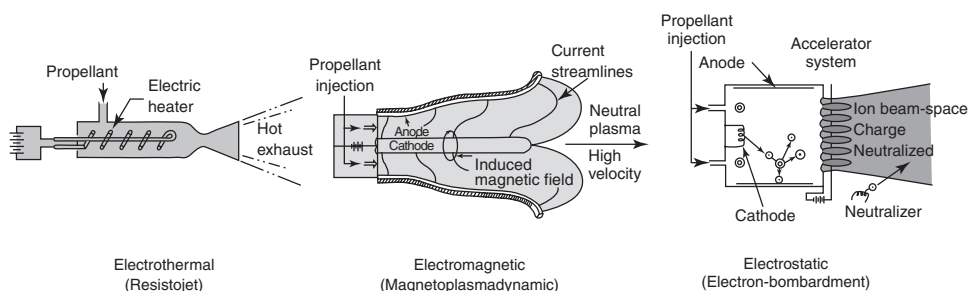


Figure 3. Electric propulsion thruster types.

Electron-Bombardment Ion Thruster. In this device, gaseous propellant is supplied to the discharge chamber and ionized by electron bombardment in a Penning-type discharge.¹ The accelerator grids extract and focus the ions in a large number of “beamlets” to form a thrust beam. The propellant is exhausted as ions, so an electron-emitting neutralizer is provided to couple electrons into the thrust beam and thereby to balance the charge expelled as ion current.

Emergence of Ion Propulsion (Electron Bombardment Thrusters)

The acceleration concepts identified in the preceding section have both advantages and disadvantages. The development status of the various technologies also differs. In addition, the technological difficulties encountered in development vary among the concepts and thus the emphasis of U.S. research and development has been governed to a large extent by considering the potential benefit versus the technological difficulties. At this time, the electron-bombardment ion thruster is the most mature and the most ground- and space-tested of the high- I_{sp} devices. Hence, it will be our major focus.

Total System Concept. The basic components of a complete propulsion system include an external source of energy (for example, sunlight incident on a solar panel), an energy-management subsystem (that supplies power to the thruster in a useful form), a propellant-supply subsystem, and a thruster that uses electrical power to accelerate the propellant and generate thrust.

Space Tests. Because of their inherently low thrust and consequent long operating requirements, extensive testing in laboratories on Earth was essential to the development of ion engines for eventual use in space. Simulating the space environment for an ion engine is much more complex than the usual space-simulation problem. Besides maintaining an excellent vacuum (of the order of 10^{-6} mmHg) in the presence of a large and continuous high-energy gas loading, in addition to the usual physical environmental factors such as solar radiation and gamma rays, an adequate way must be found to simulate the electrical environment of space. The walls and collectors must not act as sources of electrons. In space, the ion beam is injected into an infinite region that has no boundary conditions, whereas, in a vacuum chamber, the walls and collector impose artificial electrical boundary conditions at the edges and end of the ion beam that may influence the realistic interpretation of an engine's performance. Another problem concerns sputtering; even a small ion engine that operates for hundreds or thousands of hours can sputter away many tens of pounds of collector material, a difficult problem with which to contend. As difficult as these problems are, they were solved in a number of government and industrial laboratory test facilities. And, as a result, more than 350,000 hours of thruster and component-life tests have been carried out with mercury, cesium, and xenon propellants (9).

Of course, the most definitive testing of ion engines is done in space. A series of ballistic and orbital tests of ion engines have been pursued, three of the most important flight-qualified systems are described here.

¹In 1960, H.R. Kaufman, at the NASA GRC, adapted this type of discharge to an ion thruster configuration (8).

Space Electric Rocket Test (SERT I). The SERT I (Fig. 4) was launched from Wallops Island on 20 July 1964 on a four-stage Scout rocket (10). It followed a ballistic trajectory for 47 minutes and had the distinction of being the first successful flight of an ion engine. A NASA Glenn Research Center mercury electron-bombardment thruster operated 31 minutes at a thrust level of 4.5 mlbf and provided data on ion beam neutralization, thrust level, radio communication interference, and differences in performance between ground and space testing. Thrust was measured by determining changes in spacecraft spin rate. From these measurements, it was determined that complete beam neutralization was achieved in agreement with vacuum chamber tests. In addition, there was no communication interference.

Space Electric Rocket Test (SERT II). The SERT II spacecraft was launched in February 1970 (Fig. 5); it contained two 15-cm-diameter mercury electron-bombardment ion thrusters each capable of producing 30-mN thrust and a 4200-s specific impulse at an efficiency of 70% (11). The original flight objective was to demonstrate long-term space operation of ion thrusters by accomplishing operating times of three and five months, respectively, for the two thrusters. Before completing their goals, both thrusters failed due to accelerator-electrode erosion that ultimately resulted in short circuits between the high-voltage electrodes of the beam-extraction system. However, the electrical short in one thruster was eventually cleared, and steady-state thruster operation was reinitiated in January 1979.

In total, the SERT II project demonstrated long-term (i.e., 22.5 years) thruster operation in-orbit and identified a failure mode that occurs only in a zero-gravity environment. Consequently, ion thrusters, now designed, operate with the neutralizer located farther from the accelerator grid, and accelerator voltages are operated at lower values to eliminate the type of erosion that caused the failure of the SERT II thrusters.

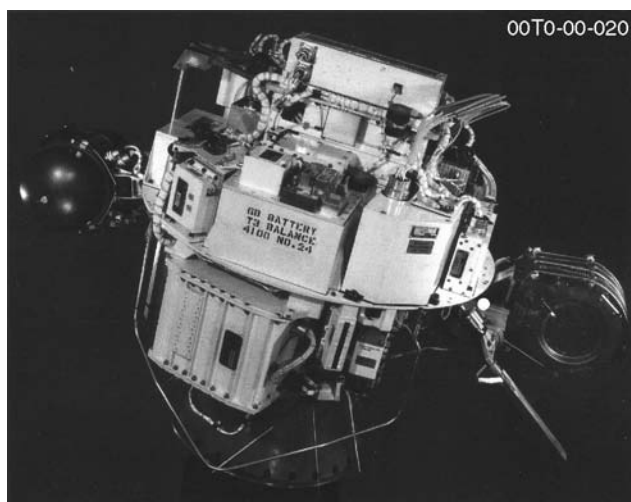


Figure 4. SERT-I ion engine flight test spacecraft (photo courtesy of NASA).

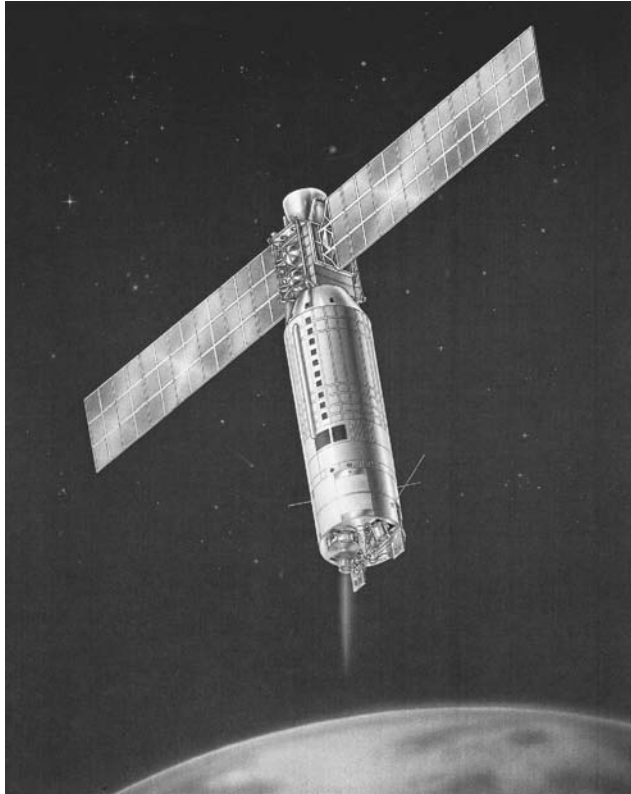


Figure 5. Space electric rocket test spacecraft in orbit (photo courtesy of NASA).

Ion Auxiliary Propulsion System (IAPS). The objective of the IAPS technology program and test flight was to develop a prototype 4.5-mN electron-bombardment thrust system and subject it to flight-qualification-level testing and long-term space operation. The test plan included operating one thruster cyclically to demonstrate the durability of the subsystem during a simulated 7-year North-South stationkeeping mission on a typical geosynchronous satellite. This test schedule implied completing 2560 cycles of 2.8 hours each, a total of 7200 hours of thrusting time. In the actual flight test, a 2-hour thruster-off period was planned between operational periods, resulting in 17 months of spacecraft operating time.

The IAPS flights hardware (12,13) consisted of two 8-cm thrust systems and a diagnostic instruments package (the two IAPS units are shown in Fig. 6). The IAPS ion thruster developed 4.5-mN thrust and operated at a specific impulse of about 2500 s at an overall efficiency of 40%. The power electronic unit contained nine independently programmable power supplies. A digital-control-interface unit provided the power-and-command interface between the user and the thrust subsystem. The digital-control interface was a signal-processing unit that transmits power and 16-bit serial digital commands. It contained a micro-processor that had all of the control algorithms to provide several set points

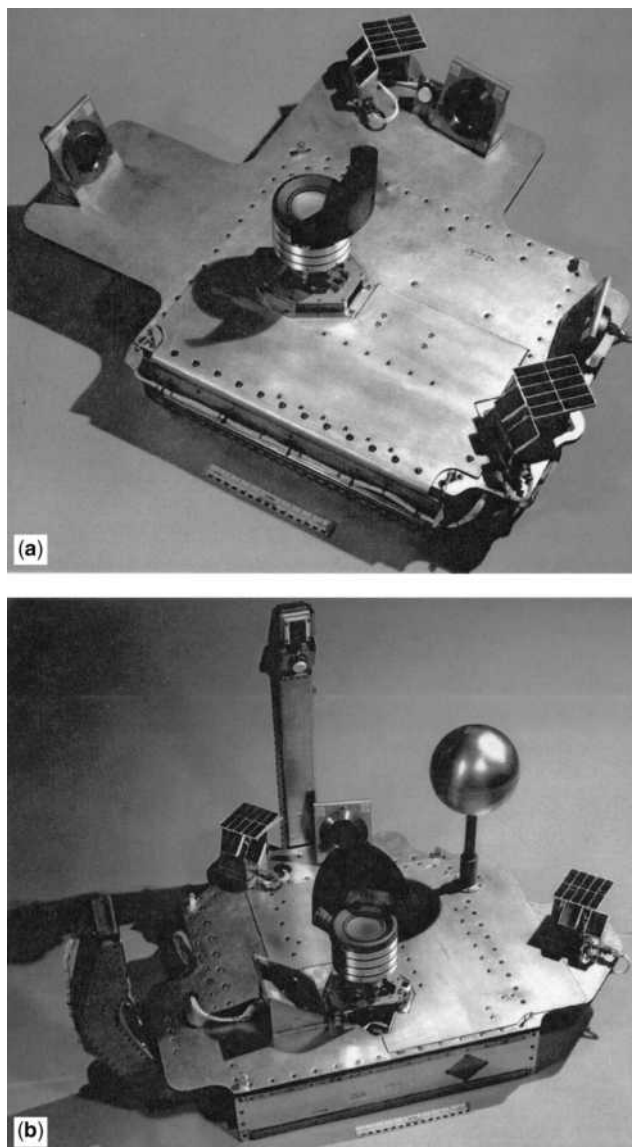


Figure 6. External components of IAPS modules include the ion thruster (shown at center) surrounded by diagnostic instruments. (a) Ram/wake module; (b) Zenith module.

for heater and vaporizer supplies, as well as proper sequencing of thruster operations.

The IAPS flight hardware was successfully subjected to extensive ground testing and was delivered to the spacecraft contractor for integration in July 1982. Spacecraft integration testing began in February 1983. However, the scheduled launch was canceled because of budgetary constraints.

Transition to Xenon. From a purely propulsive standpoint, the optimum propellant is one that has the largest atomic mass. Consequently, early

development emphasized mercury propellant and mercury ion thruster technology. However, the use of gaseous propellants that are non toxic and noncondensable has several advantages over mercury from overall system and program considerations. For cyclic thrust requirements, ion thrusters operating on gaseous propellants do not require complex thermal designs or carefully orchestrated preheating to prevent propellant condensation during thruster start-up. In addition, thrusters operated on gaseous propellants can be shutoff abruptly without an uncontrolled loss of neutral propellant during thruster cooldown after the thrusting period. In ground testing, thruster propellant flow can be measured without waiting for the thruster to reach thermal equilibrium. Because of these system and program considerations, advanced thruster development became focused on thruster operation using inert gases.

Inert-gas ion thrusters require higher ion beam current to produce a given thrust; this increase in current is inversely proportional to the ion mass (at constant specific impulse). Consequently, inert gas thrusters must produce ions more efficiently in the discharge chamber (measured in watts per beam ampere). Using advanced discharge chamber designs, thruster efficiencies have been obtained from Xe, Kr, and Ar (14) propellants that equal or exceed mercury ion thruster performance (e.g., total efficiency² = 70%) for a specific impulse in the range of 3000–6000 s.

Inert-gas ion thrusters require fewer power supplies and may be controlled by using less complex algorithms than those required for controlling mercury ion thrusters (because thruster thermal variations do not change propellant flow characteristics). Hence, since 1984, development programs have been directed toward xenon electron-bombardment ion thrusters.

Detailed Thruster Design and System Description. The xenon electron-bombardment ion thruster is now the primary contender for both auxiliary and primary propulsion applications in the near future, so we will emphasize and detail the descriptions of this thruster and its associated propellant feed and power conditioning subsystems. These descriptions will include some of the physics as well as engineering aspects of the resulting ion propulsion systems.

A schematic of the xenon ion thruster (15) is shown in Fig. 7. The thruster consists of an ionization chamber, an ion-extraction assembly, and an ion-beam neutralizer, typically, the ionization chamber is a cylinder with an electron-emitting hollow cathode (16) located at one end and an ion-extraction assembly located at the other. Xenon atoms enter the ionization chamber through the propellant plenum (about 80% of the total gas flow) and the hollow cathode (about 10% of the total flow). Electrons emitted by the hollow cathode are attracted to the cylindrical wall, or anode, which is maintained about 30 volts positive with respect to the cathode by the discharge power supply. The electrons are prevented easy access to the anode by a magnetic field, which is generated by rings of rare-earth permanent magnets (14). The energetic electrons work their way across the magnetic field lines by collisions, primarily with xenon atoms.

²Total efficiency (η_T) equals electrical efficiency (η_e = ion beam power/input power) times propellant utilization efficiency (η_m = fraction of propellant ionized and accelerated) times the square of a factor that accounts for beam divergence and multi-charged ions.

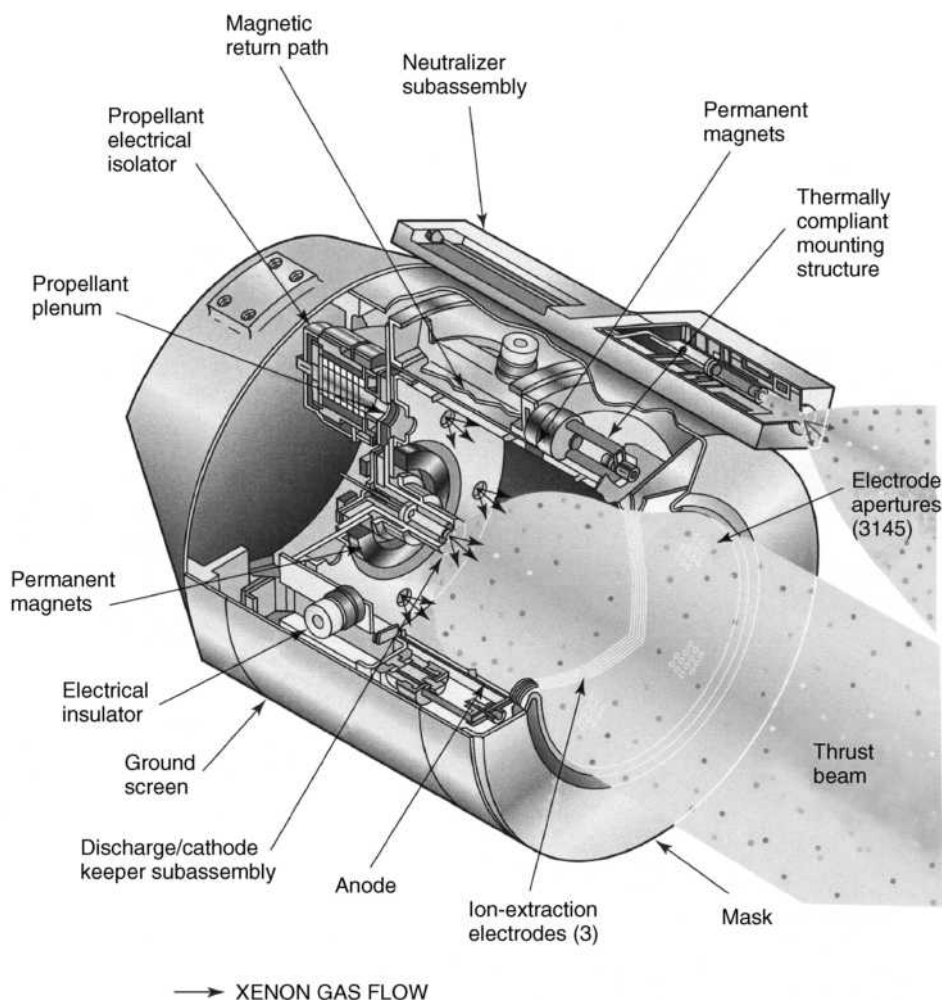


Figure 7. Schematic of xenon ion thruster.

Some of these electron-atom collisions result in ionizing the atom, and electron is released to form a positive xenon ion Xe^+ .

Under normal operating conditions, the ionization chamber is filled with plasma, in which the ion and electron densities are about 10^{17} m^{-3} and the xenon atom density is about 10^{18} m^{-3} . Ions preferentially drift toward the downstream end of the chamber at a velocity (the Bohm velocity) determined by the electron temperature, which is $\approx 5 \text{ eV}$. The gas atoms have a speed determined by the wall temperature, which is $\approx 500 \text{ K}$. As a result of the tremendous difference in their velocities, ions exit the chamber at a rate about 10 times greater than the atom loss rate; discharge propellant utilization efficiencies of 90% or more are typical of these devices.

The ion-extraction assembly consists of a pair of grids, or electrodes, that are spaced very closely ($\approx 0.5 \text{ mm}$) and contain several thousand precisely

matched apertures. The inner, or screen, grid is maintained at a positive potential of the order of 1 kV (with respect to the spacecraft) by the screen power supply. The outer, or accelerator, grid is maintained at a negative potential of the order of a few hundred volts by the accelerator power supply. As the ions approach the apertures in the screen grid, they are attracted by the negative accelerator grid and are accelerated through the sum of the potentials applied to the two grids plus the anode voltage. After the ions exit to space through the accelerator apertures, they are decelerated to essentially zero potential (space-plasma potential). As a result of this acceleration/deceleration action, the sum of the voltages applied to the anode and screen grid determines the final velocity of the ions. The negative potential of the accelerator grid also provides an electrostatic barrier that prevents the positive potential of the screen grid from accelerating electrons in the ion-beam plasma through the apertures in a direction opposite to the ion flow.

Some ion thrusters use a third electrode, or decelerator grid, located downstream of the accelerator grid. The decelerator grid establishes the precise axial location of the neutralization plane, and it allows more freedom in selecting the voltages applied to the screen and accelerator grids (i.e., selection of the accelerator/decelerator ratio). A decelerator grid also has the advantage of shielding much of the spacecraft from direct view of the accelerator grid. This shielding is important because the accelerator grid undergoes charge-exchange ion sputtering during thruster operation and is a potential source of contamination for sensitive elements of the spacecraft, such as solar arrays, thermal radiators and blankets, and optical sensors.

The ion-beam neutralizer (17) consists of a hollow cathode similar to the one that supplies electrons to the ionization chamber. The neutralizer provides electrons to match the ion-beam current and to neutralize the space charge of the positive ion beam. Current neutralization is required to prevent the spacecraft from charging negatively due to the ejection of positive charge. Space-charge neutralization is required to prevent excessive divergence of the positive ion beam due to the mutual repulsion of like charges. The neutralizer cathode consumes propellant but produces essentially no thrust. Therefore, highly gas-efficient neutralizers are necessary to achieve high specific impulse and thruster efficiency. The neutralizers in use today consume as little as 6% of the total xenon required for thruster operation.

The lines that provide xenon to the propellant plenum, discharge cathode, and neutralizer cathode contain propellant electrical isolators, such as that shown in Fig. 8. These devices allow keeping the propellant storage and control system at the spacecraft (ground) potential, whereas the thruster components may be at potentials as high as kilovolts. The isolator consists of a cylinder constructed of alumina that has overlapping shields surrounding its outer surface to prevent it from being coated by sputtered electrically conductive material, especially during ground testing. A stack of stainless steel screens and alumina spacers is contained within the body. The thickness of the spacers ensures that the voltage across adjacent screens is less than the minimum Paschen breakdown voltage for xenon. By this arrangement, the applied voltage divides linearly along the length of the isolator, and the number of individual screen/isolator elements establishes the total voltage-isolation capacity.

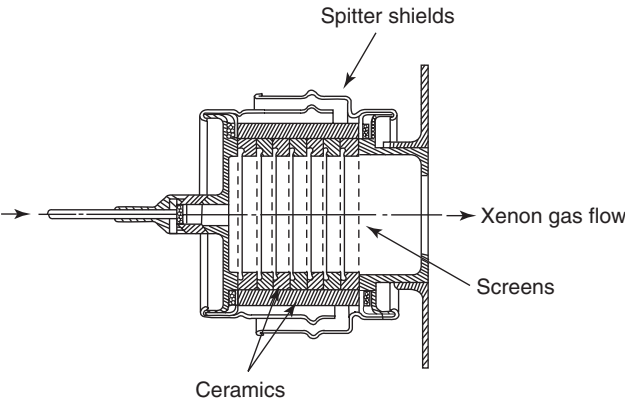


Figure 8. Propellant electrical isolator.

A power processor unit (PPU) provides the currents and voltages required to power the various components of the thruster and the controls for programming the individual power supplies for startup, steady-state operation, shut-down, and overcurrent protection. Typical power supply outputs for auxiliary and primary propulsion thrusters are listed in Table 3. The PPU receives its input power directly from the spacecraft bus and typically provides seven outputs to the thruster in the form of regulated currents or voltages.

The xenon storage and control system consists of a high-pressure tank and a flow controller. Xenon is stored at an initial pressure of about 100 atm, where its specific gravity is about 1.65. The gaseous xenon is metered into the thruster cathodes and the ionization chamber by a flow controller that may be an active or a passive system. Active flow controllers use valves and plenum chambers to maintain the individual gas flows at constant values or to control them in proportion to critical thruster electrical parameters such as the anode voltage, neutralizer keeper voltage, and beam current. Passive flow controllers use a pressure regulator and individual orifices to provide constant gas flow rates to the two cathodes and the ionization chamber. The required (constant) gas flow rates are determined during initial flight acceptance testing of a thruster.

Ground and Space Life Tests. The first long-term life test of a xenon ion thruster was completed in 1988 (18). This test was actually a wear-mechanism test

Table 3. Power Supply Output Characteristics for Auxiliary and Primary Propulsion Applications of Xenon Thrusters

Output/Application	Auxiliary propulsion	Primary propulsion
Screen	0.5 A at 750 V	3.5 A at 1200 V
Accelerator	10 mA at 200 V	100 mA at 450 V
Discharge	4 A at 40 V	21 A at 40 V
Discharge keeper	1 A at 15 V	1 A at 10 V
Discharge heater	3.6 A at 10 V	3.8 A at 12 V
Neutralizer keeper	1 A at 15 V	1 A at 10 V
Neutralizer heater	3.6 A at 10 V	3.8 A at 12 V

involving a 25-cm-diameter laboratory-model thruster that had not been designed for space application nor fabricated or assembled using the quality standards imposed on high-reliability space hardware. In spite of the level of the hardware involved, this test was successfully completed after achieving 4350 hours of accumulated ON time in 3850 ON/OFF cycles of approximately 1-h duration.

The thruster was operated at an input power of 1.3 kW and produced about 65 mN of thrust at a specific impulse of 2800 s. The hours and cycles successfully demonstrated before the test was terminated were approximately twice the operating requirements of this thruster for the intended application of N-S stationkeeping on a large spin-stabilized communications satellite.

Several additional life tests of engineering-model and flight-model xenon ion thrusters have subsequently been performed at NASA, JPL, Hughes Electronics, and elsewhere, involving thrusters that ranged in size from 10- to 30-cm beam diameter (19–21). The longest duration life tests of xenon thrusters are currently underway at Hughes Electronics and had accumulated 13,000 and 14,000 hours when this article was written. These units have 13-cm-diameter beams and operate with an input power of about 440 W to produce 18 mN of thrust at a specific impulse of 2565 s. The highest power life test is also currently underway at Hughes Electronics, and when this article was written, it had accumulated more than 1500 hours at an input power of 2 kW and an additional 3000 hours at an input power of 4.2 kW. This unit has a 25-cm-diameter beam, and at these power levels, it produces 80 or 165 mN of thrust and achieves a specific impulse of 3700 s.

Life testing of xenon ion thrusters has confirmed that operating times of 10,000 to 20,000 hours are feasible and that the primary life-limiting mechanisms are accelerator-grid erosion due to sputtering by charge-exchange ions and screen-grid erosion due to sputtering by doubly charged ions. The charge-exchange ions are formed in the ion acceleration and deceleration regions, which contain both primary beam ions and un-ionized xenon atoms. In the resonant charge-exchange reaction, an electron is exchanged between the atom and ion, creating a slow charge-exchange ion. Sputtering of the negative accelerator grid occurs when the charge-exchange ions are accelerated to it.

The plasma potential within the ionization chamber is roughly equal to the anode potential. Therefore, surfaces that are at cathode potential, such as the screen grid, are continuously bombarded by ions that fall through the anode-to-cathode voltage. Doubly charged ions are formed within the ionization chamber, primarily by a multistep ionization process in which a singly charged ion is bombarded by an electron. The doubly charged ions are particularly damaging because they strike the screen grid at energies equal to twice that of a singly charged ion. The sputter yield of typical grid materials, such as molybdenum, increases exponentially with ion energy, and models of the screen grid sputtering show that nearly all the sputtering in ion thrusters is caused by doubly charged ions (22,23).

Other components that could limit the lifetime of ion thrusters include the cathodes and associated heaters. However, cathode-life tests (24) have successfully achieved 28,000 hours of cyclic operation on xenon, and heater tests have demonstrated more than 45,000 ON/OFF cycles.

Xenon thrusters have been flight-qualified by rigorous ground testing in a variety of sizes, including beam diameters of 10, 13, 25, and 30 cm, and the

13-, 25-, and 30-cm-diameter thrusters have also been space-qualified on commercial or scientific spacecraft. Hughes Electronics has launched 10 communications satellites that use their 13-cm and 25-cm xenon ion propulsion subsystems (XIPS) for stationkeeping, orbit raising, momentum dumping, and attitude control. When this article was written, more than 7000 hours of operation had been accumulated on the 40 XIPS thrusters that have been launched, and more than 57,000 hours of acceptance, qualification, and life testing had been accumulated in ground testing.

Manufacturing Status. For xenon ion propulsion systems to be widely applied to satellite and planetary/interplanetary missions, an industrial manufacturing capability must be developed. The beginning of this production capability has already been initiated with significant effect. For example, Hughes Electronics already maintains a production-line manufacturing capability for both 13- and 25-cm thrusters, and their current manufacturing and test capacity is about six flight units per month.

In addition to manufacturing satellite control propulsion systems, producing new primary propulsion systems is already underway. This latter production need was initiated by NASA's Deep Space 1 (DS-1) mission, which used a 30-cm-diameter thruster and PPU (called the NSTAR system, NASA Solar Electric Propulsion Technology Applications Readiness) for its primary propulsion (25,26). The DS-1 spacecraft, shown in an artist's conception (Fig. 9), propelled by the NSTAR ion-propulsion system, covered a distance of more than 150 million miles on a trajectory that featured flybys of the asteroid Braille and the comets Wilson-Harrington and Borrelly. The successful completion of the DS-1 mission qualified solar electric propulsion for use on future NASA missions, both Earth-orbital and deep space. As a result, an industrial manufacturing capability for the NSTAR thrust system has been established at Hughes Electronics.



Figure 9. Deep Space 1 (DS-1) spacecraft (artist's concept, courtesy of NASA).

Applications

Space-mission interests and opportunities continue to expand as they reflect the capability of space launchers, new mission requirements, space policies, and economies of scale. Future space missions will become increasingly demanding in energy and duration. Therefore, they will profit more significantly from using ion propulsion.

These missions range from near-Earth vehicle precision position control and orbital change to interplanetary missions. During the past decades, the specific payoff of ion propulsion has been identified for each of these broad ranges of applications; a sampling of these results are reviewed in the following paragraphs with more specific descriptions.

Synchronous Satellites (Auxiliary Propulsion). Two sources of perturbing forces dictate the control system requirements for maintaining a satellite stationary in a 24-hour orbit: the gravitational attraction of the Sun and Moon and the triaxiality of Earth (e.g., see Fig. 10). The magnitude and nature of these perturbations determine the design and mode of operation of the ion propulsion stationkeeping system.

However, because correction for the triaxial perturbation (27) requires thrust in an E-W direction or perpendicular to that of the N-S solar-lunar effect correction, either separate or gimballed ion thrusters would be required. Because of the low energy required, this E-W correction under normal conditions would not be considered a highly useful application of ion propulsion. Therefore, let us concentrate on N-S satellite stationkeeping.

General Considerations. Of all celestial bodies, only the Sun and Moon produce significant perturbing effects on an Earth satellite at synchronous

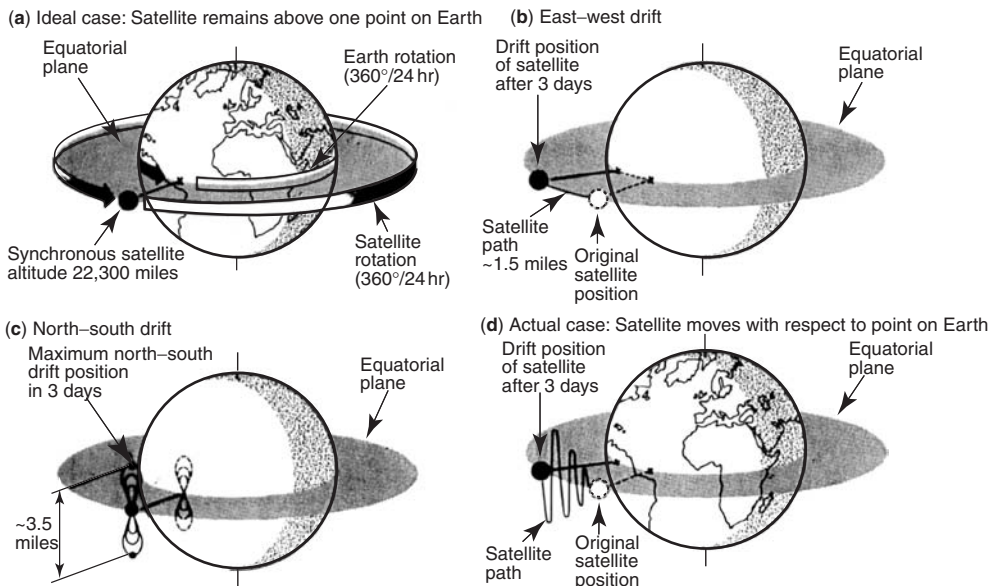


Figure 10. Synchronous satellite perturbations.

altitude. The force of attraction from these two bodies can be divided into three components: two, the radial and tangential, lie in the equatorial plane, and the third lies normal to this plane. The radial and tangential components produce cyclic oscillations in satellite radius and longitude. It can be shown that these oscillations have a small maximum amplitude throughout each satellite orbital period and that at the end of that time, both the change and rate of change of satellite radius and longitude will have assumed their initial zero values. Hence, no corrective thrust is required to counteract the radial or tangential force components.

The component of force normal to the equatorial plane (for synchronous satellites, this will be referred to as the north-south or "Z" component) will, however, cause an oscillation that results in increasing the satellite's orbital inclination at an initial rate of 1° per year (28). The amplitude of this oscillation will grow to a maximum of 20° in about 40 years. If the satellite is to remain stationary in the equatorial plane, it is necessary to correct this perturbation. The optimum mode of correction and the magnitude of thrust can be determined by solving the appropriate differential equation while taking corrective thrust into consideration.

The solution to this equation is shown in graphical form in Fig. 11 (29), where the uncorrected Z-oscillation is given for a 48-day interval during which the Sun passed through a point of maximum perturbation. The maximum Moon effect occurs every 13.6 days. Figure 11c illustrates the total solar-lunar effect on a satellite's orbital inclination for a representative 48-day period. The inclination at the end of this time is 0.16° . The effects of solar and lunar perturbations are shown separately for the same period in Figs. 11a and 11b, respectively. The total change per year in orbital inclination can be determined from these curves. The Sun's contribution calculated from the data is 0.30° , whereas that of the moon is 0.64° , a total of 0.94° per year. From this result, the minimum daily amount of stationkeeping propulsion can be calculated.

To counteract the change in inclination and maintain the satellite in the equatorial plane, velocity must be added to the satellite in a direction normal to the orbital plane. Because of the gyroscopic nature of the satellite in its orbit, the orbit will precess about an axis directed to the point of applied torque and normal to the thrust. If the satellite is to be precessed into the equatorial plane, corrective thrust is most efficient when applied impulsively at the nodal points, (i.e., at satellite crossings of the equatorial plane). (However, it will be shown later that impulsive firing is not the optimum, or even the necessary, thrust mode for an ion propulsion system). These crossings occur twice daily for a 24-hour satellite.

As an example, assume a thrust mode of two nodal firings per day, alternating in a north-south direction, depending on whether the satellite is crossing an ascending or descending node. The total impulse required each day to correct the total solar-lunar perturbations (0.94° per year) on a 550-lb satellite is 7.8 lb-s. If, for example, 1.5-mlb stationkeeping ion engines are employed, thrust must be applied for 43.5 minutes per firing. Because the required thrust time, when correcting twice daily at the nodes, can be determined analytically, this correcting mode (now our reference) can be used to compare the effectiveness of other thrusting sequences. Figure 12 (29) demonstrates clearly that each of the five

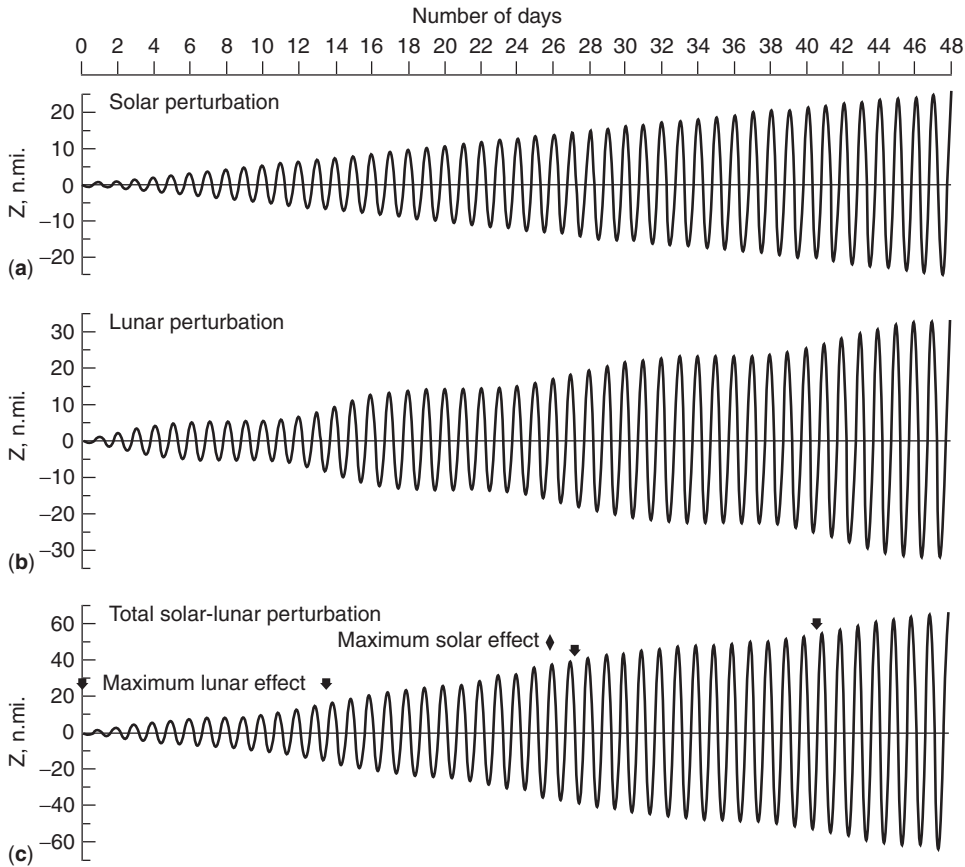


Figure 11. Typical daily longitudinal variations of 24-hour satellites due to solar and lunar perturbations.

examples of thruster firing duration provides approximately the same control of inclination variation. This conclusion gives some flexibility to the satellite vehicle designers.

Note that Fig. 12 shows variations in satellite orbital inclination at maximum solar-lunar perturbation times. Hence, continuing the indicated thrust schedule will keep the satellite within the bounds shown.

Present Stationkeeping Systems. A typical layout³ of a contemporary string-redundant ion propulsion system for north-south stationkeeping maneuvers on geosynchronous satellites is shown in Fig. 13. A single north and south thruster can provide a $\Delta V \approx 750$ m/s required to maintain the orbit of a GEO satellite in the Earth's equatorial plane for a 15-year period. Two thrusters are powered by a common PPU, minimizing the overall mass of the propulsion system (30). Using this arrangement, voltages are applied simultaneously to the

³Actually, the layout shown here is representative of the 13-cm XIPS system that Hughes Electronics has implemented on its HS-601HP communications satellite product line.

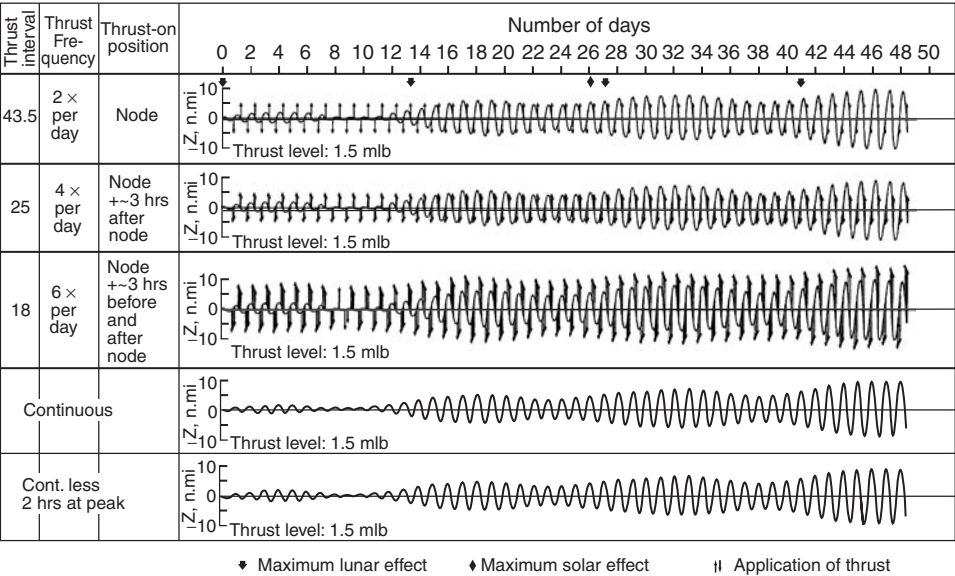


Figure 12. Latitudinal control using various thrust sequences.

various electrodes (i.e., screen, accelerator, decelerator, anode, and keepers) of the north and south primary thrusters (N1 and S1). However, cathode heater power and xenon gas flow is provided only to the thruster selected for the propulsive maneuver (see the north thruster at the ascending node). Twelve hours later, when the satellite approaches the other nodal crossing, the same PPU applies voltages to the electrodes of the primary thrusters N1 and S1, but gas flow and heater power are provided only to the selected thruster (see the south thruster at the descending node). Full redundancy is achieved by using the other PPU and the secondary thrusters N2 and S2. The two thruster pairs shown in Fig. 13 (a primary and a secondary) are mounted on a motorized platform that can be gimballed in two axes to maintain the thrust vector pointing through the spacecraft's center of mass.

The xenon storage and control system typically consists of two xenon tanks isolated by a squib valve and a dual-string arrangement of a squib valve, two-stage pressure regulator, and series-redundant latching valves. The string approach illustrated in Fig. 13 typically uses the primary thrusters for the first half of the mission and the secondary thruster pair for the second half of the mission. If a regulator fails in either leg of the system, the squib valve that normally isolates the two legs is fired so that both tanks can provide xenon to the remaining usable thrusters.

As described earlier north-south stationkeeping is accomplished by firing a north thruster daily for a period of approximately 5 hours centered about the ascending node. Twelve hours later, the south thruster is fired for the same duration, centered about the descending node. Because the thrusters are typically aimed away from Earth, firing them introduces a radial component of thrust that would normally make the orbit undesirably eccentric. However, this twice-per-day thrusting strategy automatically removes the eccentricity in any 24-hour

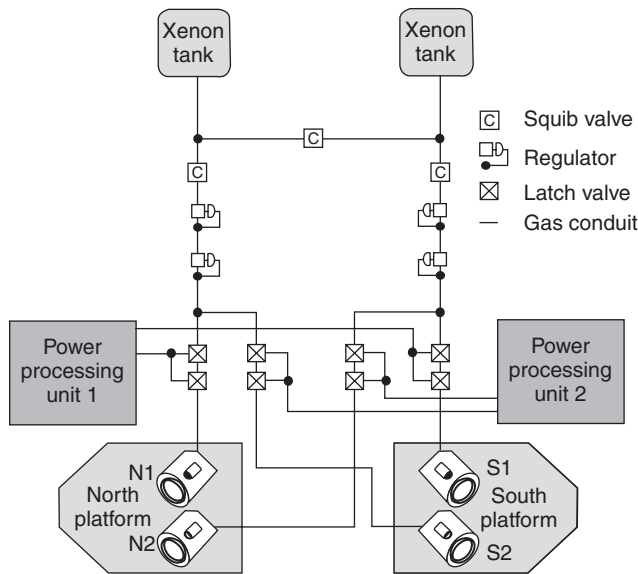


Figure 13. Layout of a modern xenon ion propulsion subsystem.

period. The propulsive maneuvers are directed by the spacecraft's computer, which follows program commands uploaded to the satellite every two weeks.

Interplanetary/Planetary Spacecraft (Prime Propulsion). Figure 14 represents the propulsive requirement for a number of interplanetary/planetary missions as a function of distance from the Sun. The curve labeled minimum energy transfers (Hohmann) describes the velocity increment required for a minimum flyby of a particular body at its position relative to the Sun. These Hohmann transfers for missions to Jupiter and beyond often require prohibitively long flight times.

The horizontal line, labeled chemical limit, describes the maximum velocity-increment, capability assuming chemical propulsion systems and a single Space-Shuttle launch. The velocity increments required for a number of missions are overlaid on this figure and expressed both for a chemical-propulsion mode and for an electric-propulsion mode. (The velocity increments required for chemical- and electric-propulsion modes differ due to the gravity loss terms for continuous thrusting during long times). A number of missions are shown that are well within the capabilities of solar-electric propulsion.

Modularization. Relatively speaking, high thrust cannot be provided by a single ion thruster because of technological problems associated with extremely large diameters of discharge chamber and ion optics components. Therefore, modularization of the total thruster system (e.g., see Fig. 15) is required (31). Modularization, however, in a positive sense, leads to two critical propulsion system design advantages: 1) the possibility of increased system reliability through redundancy and 2) the ability to power-match the solar panel power source to the propulsion system's electrical load.

System Reliability. An important question to be answered before system design can begin is the total number of modules to be employed in building up to

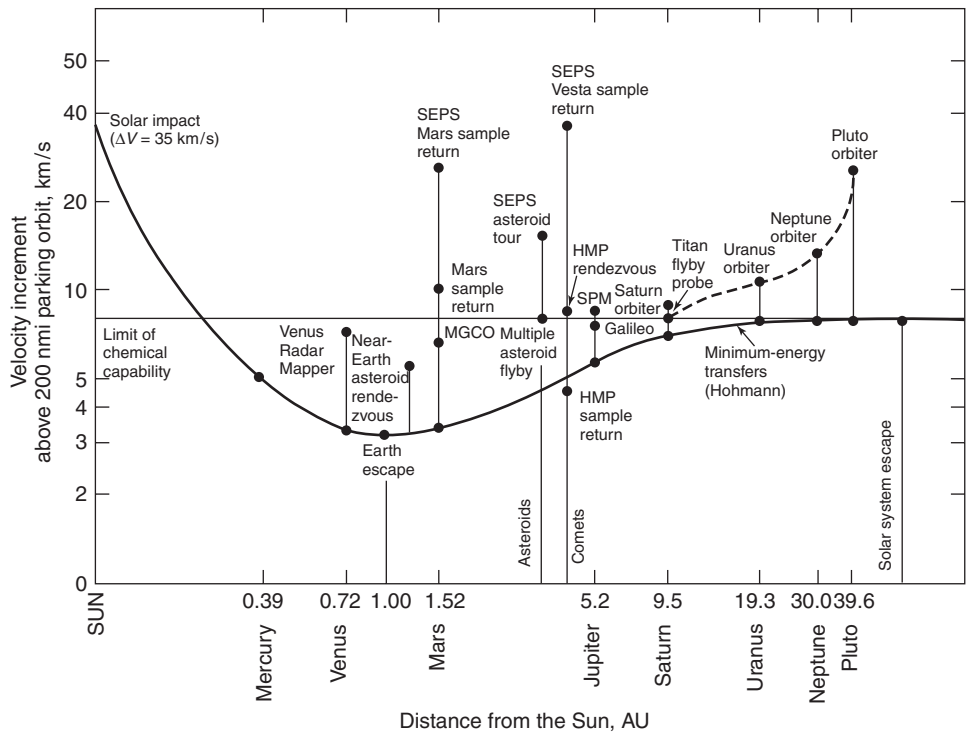


Figure 14. Interplanetary mission propulsion requirements.

a high-power electric propulsion system (32). Several important considerations affect the answer to this question; the foremost is system reliability.

As is well known, establishing component reliability figures at high confidence levels requires costly and time-consuming test programs. It is possible, however, to build up system reliability through redundancy techniques even though component reliability is somewhat low or possibly not even established. Redundancy, whether series, parallel, or standby, will increase propulsion system weight. Therefore, it is always desirable, to determine the method by which the requisite reliability can be obtained by adding a minimum to system weight.

Standby redundancy proves especially applicable in the engine and power conditioning systems, provided that each of these systems is composed of a number of identical modules. In addition to incorporating initial standby units in a modularized ion propulsion system, it will be shown in the next section that, because of the decrease in available power and increase in output voltage of the solar panel in deep-space missions, engine and power conditioning modules will be shut down during the course of the mission. In each case, the modules and their respective subsystems will be designed so that disconnected modules can be reconnected. From the view point of reliability, the shutdown modules can be considered standbys.

Power Matching. As mentioned in the previous section, a major challenge in designing a solar electric propulsion system is the use of the maximum power available from the solar panels. The thruster and power conditioning systems

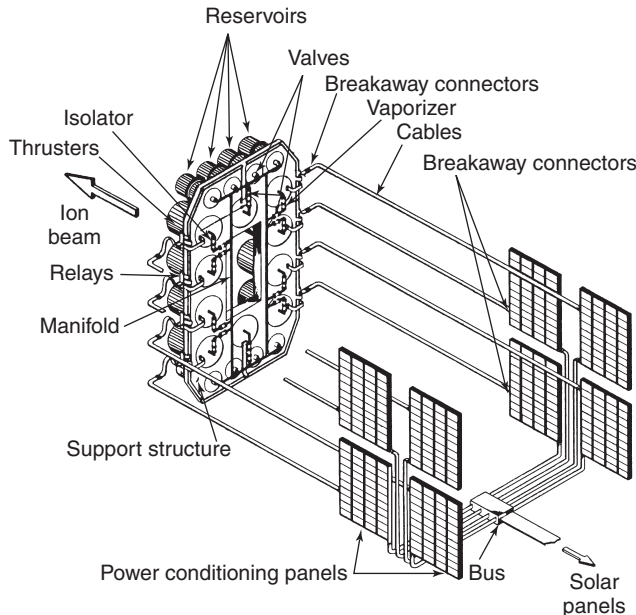


Figure 15. Conceptual drawing of a 48-kW ion propulsion system.

must be designed and programmed to provide the proper load at all times to the constantly varying power and voltage outputs of the solar array. Figure 16 shows a family of normalized nonlinear I - V characteristics for a solar panel on a trajectory directed toward the outer reaches of the solar system. This example assumes a modularized ion engine system that begins the mission by using 16 modules at full thrust and, as solar panel power decreases, engines are switched off at the appropriate times to maximize power transfer from panel to engine load (33).

At any particular time during the example mission, the system must operate at a point located somewhere on the I - V curve, which corresponds to that time. For each I - V curve, a point $[I(t), V(t)]$ can be found at which the power ($P = IV$) is a maximum. To operate at the point $[I(t), V(t)]$ and thereby derive the maximum power possible from the solar panel at time t , the resistance of the load R_L must be matched to that of the source [i.e., $R_L = R(t) = V(t)/I(t)$].

Maximum power is transferred from the solar cell source when the load resistance is equal to the source resistance of the solar cell supply. Because of the motion of the solar cell array away from the Sun, its I - V characteristics will vary with time during the mission. Therefore, its source resistance R will be a function of time and actually will increase during flight away from the Sun. If the load is thought of as a constant-current device, its resistance can be increased to follow $R(t)$ in two ways: (1) the propellant flow rate (and, thus, the beam current) can be decreased continuously so that the $R(t)$ curve is followed identically, or (2) single engines of the modularized thruster system can be switched off at appropriate times to maintain the load resistance close to $R(t)$. If \dot{M}_p (propellant flow rate) is adjusted continuously during the mission, the load resistance will always be matched to $R(t)$, and maximum power transfer will be maintained. In this case, the

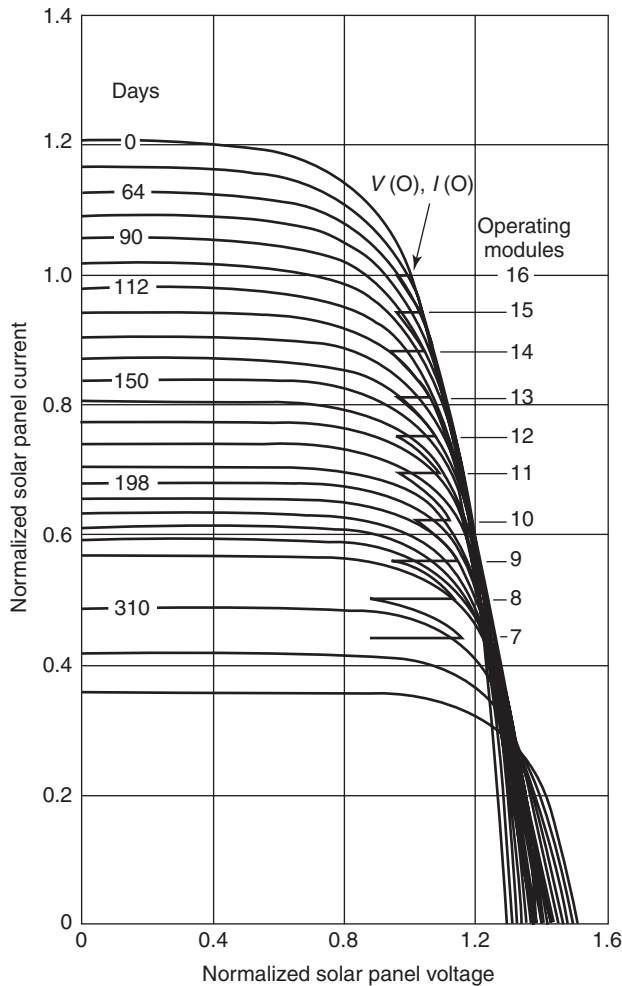


Figure 16. Normalized solar panel I – V characteristics for example trajectory away from the Sun.

current in each engine system must be capable of varying 60% during the mission, and the engine voltage will have increased 17% at the end of the mission. Operation across this range of beam current and voltage is, at best, undesirable from the standpoint of ion engine design. However, switching engines and power conditioning inverter modules off and returning all of the remaining modules to full thrust conditions can reduce these required operating ranges to 10% or less. This level of variation is well within the capability of present day ion propulsion systems.

Early Mission and Spacecraft Designs. The initial (e.g., late 1960s to 1970s) design philosophy adopted for the general arrangements of the solar-powered electric-propulsion spacecraft was one in which designs 1) were formulated to be adaptable to as wide a variety of interplanetary missions as possible for a specific launch vehicle and 2) could scaled to be compatible with a variety of existing NASA family launch vehicles. This concept was achievable by modular

design of the electric-propulsion system that allowed using of the same thruster and power-conditioning system modular designs, in varying numbers, to accomplish a variety of missions with various launch vehicles.

During these early studies, for example, two baseline conceptual spacecraft designs were primarily considered. One was a spacecraft suitable for a Mars orbiter and lander mission launched by Titan III-M with a hammerhead shroud, and the other was a universal spacecraft for a Jupiter flyby or out-of-the-ecliptic mission launched by an Atlas SLV-3C/Centaur with a newly modified shroud (34). The power level required at 1 AU for the ion propulsion systems for the two baseline designs was established as 9.6 kW for the Mars mission using the Titan III-M launch vehicle and 17 kW for the Jupiter flyby and out-of-the-ecliptic missions using the Atlas SLV-3C/Centaur launch vehicle.

Even though technological advancements and identification of new spacecraft concepts will continue, the conceptual bases of these designs still demonstrate the propulsion system elements for developing future high-power solar-electric propulsion of interplanetary/planetary spacecraft. They also demonstrate and emphasize the intimate design aspects of solar-electric propulsion and its host spacecraft.

Mars Mission Spacecraft. A conceptual spacecraft design was developed for the 9.6-kW power level Mars orbiter and lander mission using the Titan III-M hammerhead shroud configuration and a folding modular-type solar array (see Fig. 17). The solar panels are deployed in a plane normal to the spacecraft's longitudinal axis into the four quadrants at the base of the spacecraft bus. A total of 20 rectangular subpanels of equal size (8×7 feet) are used to provide a gross area of 1120 square feet. To provide adequate thruster exhaust beam clearance for the solar array, the outboard auxiliary subpanels have been hinged to deploy asymmetrically, eliminating impingement of the exhaust flux on the panels. The main hinge attachment of the solar array to the spacecraft is at the base of the spacecraft bus. The spacecraft longitudinal axis is along the sun-probe line, and the solar cells are placed on the surface away from the spacecraft bus. In this manner, the entire spacecraft bus remains in the shade, thereby relaxing thermal control requirements. The solar array subpanels are purposely limited to 8 feet in length, so that they do not extend forward (in the stowed configuration) beyond the cylindrical portion of the hammerhead shroud. By preserving shroud stowage volume forward of the cylindrical section, various lander shapes, whose maximum diameters are up to approximately 11 feet, may be accommodated. The Apollo-type lander capsule depicted here is 8 feet in diameter.

In the four-quadrant solar array configuration, the panels are stowed in a box-like fashion around the cylindrical spacecraft bus, limiting the bus to a 7-foot-diameter cross section. The spacecraft's cylindrical structure is locally squared off to accommodate the electric engine array. Placement of the engine array external to the basic spacecraft frame and shell permits greater freedom for thruster array translational motion and thrust vector angular motions, which may be desirable for some missions.

Power-conditioning modules are attached to the spacecraft bus on each side of the engine and are in shade cast by the primary solar panels. The high-gain antenna is stowed within the base of the spacecraft bus and gimbal-mounted to satisfy clock and cone angle requirements for Earth communications. A 7-foot-diameter

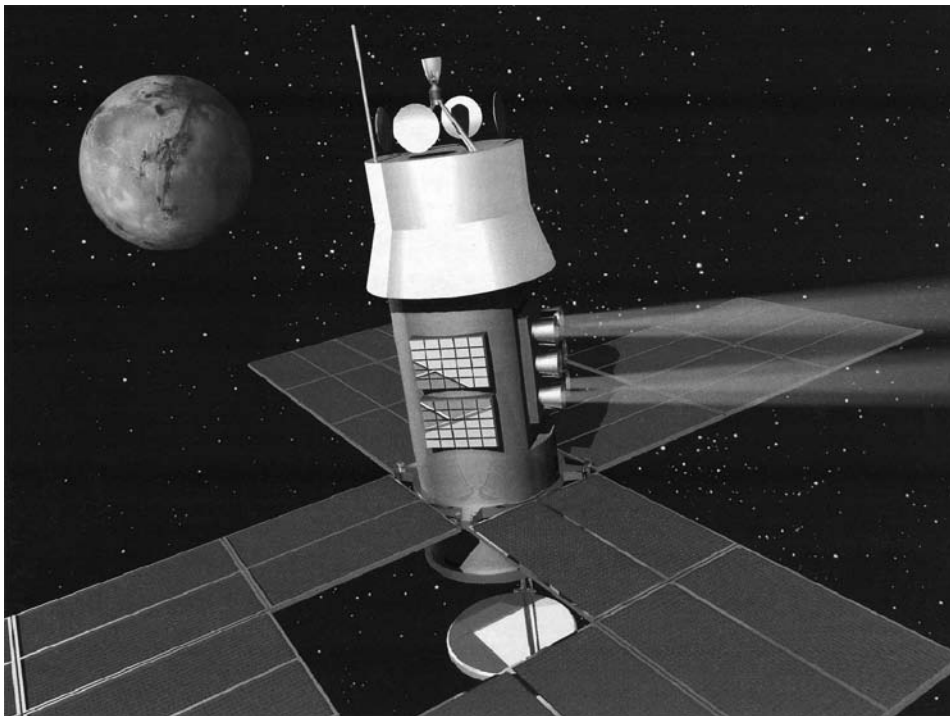


Figure 17. Titan III-M SEP Spacecraft in transit to Mars.

planar array antenna is indicated, although a 1-foot increase in diameter may be incorporated within the constraints imposed by the launch-vehicle geometry.

Major subsystems such as the nitrogen gas tanks for attitude control, bus or body-fixed scientific payload, communication electronics, and guidance and control equipment are housed within the spacecraft bus frame. The temperature of the spacecraft is controlled by two sensor-actuated louver systems.

The spacecraft bus, including the lander capsule, is to be retromaneuvered into the orbit of Mars by a liquid chemical rocket. All spacecraft configurations designed for missions presented here assume that a liquid bipropellant retro-rocket subsystem is used. The systems have been sized for a specific impulse of 315 s for a liquid bipropellant thruster that uses MMH fuel and N_2O_4 oxidizer. The retrorocket and tankage are mounted so that the thrust vector is directed along the spacecraft's longitudinal axis. The four solar panels beyond the out-board hinge of the primary panels are separated and ejected radially away from the spacecraft before the retromaneuver.

The planet-scanning scientific equipment is mounted to deploy into the space between two adjacent solar panel assemblies on the opposite side of the thruster array to obtain maximum unobstructed viewing. Two low-gain omnidirectional antennae are indicated. They are located on the spacecraft where they do not depend on deploying any other subsystem before they operate normally. The primary low-gain antenna is deployed out of the base after spacecraft/launch-vehicle separation.

The attitude control nitrogen storage tanks are mounted symmetrically about the center of mass of the deployed spacecraft configuration so that they do not contribute to center of gravity displacement as gas is consumed throughout the mission. The attitude control thrusters are located at the outer edge of the four innermost solar panels. The retrorocket fuel and oxidizer tanks are located symmetrically about the retrothrust chamber.

Jupiter Mission Spacecraft. The ion propulsion module used for the Mars spacecraft can be adapted to a mission to Jupiter. The selected mission requires an ion engine system rated at 17kW at 1AU. The Mars spacecraft previously considered imposed major design considerations dictated by the mission profile. The same is true of the Jupiter mission. Of primary concern was the large dynamic thermal range introduced by the large temperature range encountered on a Jupiter mission trajectory. For reasons of thermal control, the spacecraft bus was placed on the shadow side of the solar array plane. Maximizing the size of the high-gain antenna (6-foot-diameter parabolic dish) negated the feasibility of stowage within the base of the spacecraft due to the space limitations imposed by the package envelope of the solar array and the 57-inch launch vehicle interface. In lieu of a circular parabolic dish, a highly elliptical dish is also considered and stowed longitudinally within the spacecraft so it can be deployed out of the base. This design was rejected due to the articulation required for deployment, plus the large volume required for stowage.

The only space solely restricted by the launch-vehicle shroud envelope that is void of spacecraft interface considerations exists at the top of the spacecraft. The main solar panel hinges and high-gain antenna are placed at the top of the spacecraft bus, thereby satisfying both thermal control and communications considerations and eliminating the requirement that the high-gain antenna be articulately deployed using a mast and associated mast-mounted pointing actuators. Placing the power-conditioning modules remotely from the solar panels significantly reduces thermal coupling.

Adequate area exists in the base structure of the spacecraft to accommodate two of the six power-conditioning panels with an unobstructed view for radiating to space. For spacecraft mass balance and thermal isolation, the bulk of the spacecraft electronics and scientific payload is situated in a compartment adjacent to the power-conditioning equipment. The equipment in this compartment is thermally controlled by an independently active louver system.

Two low-gain antennas are deployed from the spacecraft bus, one to the normal sun-side position and one to the shadow side from the base of the spacecraft. Deploying either low-gain antenna is not contingent on deploying any other spacecraft subsystem. A medium-gain horn is mounted on top of the feed-supporting tripod of the high-gain dish antenna.

The ion propulsion system is composed of eight thrusters, six propellant tanks, and associated control and switching components. The thruster array is placed so that the exhaust beam is directed between two adjacent solar panel assemblies. The engine cluster is fixed so that its thrust vector is directed 90° to the Sun line. To accommodate the exhaust beam divergence of the ion thrusters without exhaust impinging on the outboard solar panels, the auxiliary subpanels are, as in the spacecraft previously discussed, asymmetrically hinged to provide

maximum clearance in the thruster quadrant. To maintain symmetrical area distribution, the panels on the side opposite the engine are arranged likewise.

The major portion of the spacecraft base is enclosed for thermal control. The central and upper sections of the spacecraft are purposely void of equipment because the most effective location for compensating for the weight of solar panels is at the base (bottom end) of the spacecraft. The attitude control nitrogen gas tank is mounted within the equipment enclosure, and the attitude control thrusters are mounted on the outboard edge of each of the four solar panel assemblies.

A truss structure extends upward to provide a square framework at the top for the main hinge attachment of the four solar panel assemblies. A light tripod supports the high-gain antenna and pointing actuators on top of the solar panel hinge frame. The truss structure in the vicinity of the engine cluster is constructed to allow longitudinal and lateral translation of the entire engine system.

Out-of-Ecliptic Probe. The out-of-ecliptic probe uses the same spacecraft (with minor modifications) as that considered for the Jupiter flyby mission. Although thermal considerations and communications requirements for an out-of-ecliptic probe do not pose design constraints as severe as those for the Jupiter flyby spacecraft, the spacecraft arrangement developed for the Jupiter mission is well suited to serve as an out-of-ecliptic probe with only minor modifications. For communications, only low-gain antennae are required. The sun-side low-gain antenna need not be deployed as on the Jupiter spacecraft, but may be permanently mounted on top of the spacecraft, simplifying the antenna mount design. For an out-of-ecliptic spacecraft, the space previously used by the high-gain antenna for a Jupiter mission may be used to accommodate approximately 25 square feet of fixed solar cell panels.

Present Deep Space Systems. NASA's NSTAR ion propulsion system consists of a 30-cm-diameter thruster, PPU, digital control interface unit (DCIU), and xenon feed system (XFS). Responding to commands received from the DCIU, the PPU can operate the thruster at any one of 16 power (thrust) levels, ranging from 500 W to 2.3 kW. (producing 19.5 to 92 mN of thrust). This wide range of power throttling enables the NSTAR thrust system to match its performance to the power available from the host spacecraft. To optimize the thruster performance across such a wide range of power throttling, the XFS adjusts the xenon flow rate to the cathodes and the ionization chamber to match the values stored in the DCIU memory. This feature adds some complexity to the XFS, but it also ensures that the thruster operates at the proper conditions to maximize performance and lifetime across the entire power range.

Conclusion

To the young scientists and engineers interested in the technological progression of humankind, we would like to use our experience in developing ion propulsion to make a point. During the 40-year research and development and eventual application of ion propulsion, there were many doubters, critics, and cynics who made known their disbelief in the efficacy of this technology! They expressed their disbelief by conjuring up the phrase, Ion propulsion is the technology of the future and always will be! Our response now is the obvious statement,

Perseverance pays off! A corollary: If you believe in what you are doing and in yourself, do not give up even in the midst of strong criticism by your detractors!

Technological Spin-offs. Ion thrusters and their associated expellant and electronics subsystems have been under development for more than four decades. During this time, numerous approaches to creating and accelerating ions for spacecraft propulsion have been investigated. Obviously, during these many years and substantial government and industrial expenditures, an extensive technological base has been established. In fact, during the evolution of these basic ion thruster technologies, the development of ion and plasma devices for a number of important nonpropulsive applications was initiated. Among these applications are some of the most advanced and critical modern semiconductor device fabricating techniques such as ion implantation and ion beam sputtering. In addition, and also of importance, are the following: (1) the advancement of secondary ion mass spectrometry, (2) the creation of spacecraft potential control, (3) the invention of a new family of power converter and switch tubes, and (4) the development of effective neutral beam injectors for fusion reactors. All of these ion propulsion technological spin-offs are more fully described in Reference 36, along with many additional specific references.

ACKNOWLEDGMENTS

The authors acknowledge the valuable contributions of Bryan M. Grossman and the Hughes Space and Communications and HRL Laboratories Information Services Departments, especially Mr. Allan Kung and Mr. W.J. Zepeda, to the production of the original draft of this manuscript.

BIBLIOGRAPHY

1. Oberth, H. *Wege zur Raumschiffahrt*. R. Oldenbourg KG, Munich, 1929.
2. Stuhlinger, E.A. *Ion Propulsion for Space Flight*. McGraw-Hill, New York, 1964.
3. Currie, M.R., and J.H. Molitor. Ion Propulsion, a Key to Space Exploration. *IEEE Student J.* 1-14 (January 1968).
4. Current Voyager Baseline Design Flight Spacecraft Weight Estimate. Jet Propulsion Laboratory, Interoffice Memo 292-2763, September 1965.
5. Saltz, L.E., and D.L. Emmons. High temperature augmented monopropellant hydrazine thruster. *JANNAF Propulsion Conf.* Monterey, CA, February 1983.
6. Jahn, R.G. *Physics of Electric Propulsion*. McGraw-Hill, New York, 1968.
7. Brewer, G.R. *Ion Propulsion*. Gordon and Breach, New York, 1970.
8. Kaufman, H.R., and P.D. Reader. Electrostatic Propulsion. D.B. Langmuir, E. Stuhlinger, and J.M. Sellen (eds). Academic Press, New York, 1961.
9. Beattie, J.R., and J.N. Matossian. Mercury Ion Thruster Technology. NASA CR-174974, Hughes Research Laboratories, Malibu, CA, 1999.
10. Molitor, J.H. Ion propulsion flight experience, life tests, and reliability Estimates. *AIAA Paper No. 73-1256*, November 1973.
11. Kerslake, W.R., et al. *J. Spacecraft Rockets* 8 (3): 213-224 (1971).
12. Hudson, W.R., et al. Electric propulsion research and technology in the United States. *AIAA Paper No. 82-1867*, November 1976.

13. Molitor, J.H. Considerations on technologies and missions for non-chemical propulsion. *AIAA Paper No. IAF-83-400*, October 1983.
14. Sovey, J.S. Improved ion containment using a ring-cusp ion thruster. *AIAA Paper No. 82-1928*, November 1982.
15. Beattie, J.R., et al. *J. Propulsion Power* 5 (4): 438–444 (1989).
16. Siegfried, D.E., and P.J. Wilbur. *AIAA J.* 22: 1405–1412 (1984).
17. Rawlin, V.K., and E.V. Pawlik. *J. Spacecraft Rockets* 5 (7): 814–820 (1968).
18. Beattie, J.R., et al. *J. Propulsion Power* 6 (2): 145–150 (1990).
19. Patterson, M.J., et al. 2.3 kW ion thruster wear test. *AIAA Paper No. 95-2516*, June 1995.
20. Polk, J.P., et al. An overview of the results from an 8200 hour wear test of the NSTAR ion thruster. *AIAA Paper No. 99-2446*, June 1999.
21. Shimada, S., et al. Ion thruster endurance test using development model thruster for ETS-VI. *IEPC Paper No. 93-169*, September 1993.
22. Beattie, J.R. A model for predicting the wearout lifetime of the LeRC/Hughes 30-cm mercury ion thruster. In R.C. Finke (ed.), *Electric Propulsion and Its Applications to Space Missions*, Vol. 79, Progress in Astronautics and Aeronautics. 1981.
23. Mantenieks, M.A., and V.K. Rawlin. Sputtering in mercury ion thrusters. In R.C. Finke (ed.), *Electric Propulsion and Its Applications to Space Missions*, Vol. 79, Progress in Astronautics and Aeronautics, 1981.
24. Sarver-Verhey, T.R. Destructive evaluation of a xenon hollow cathode after a 28,000 hour life test. *AIAA Paper No. 98-3482*, June 1998.
25. Hamley, J.A., et al. The design and performance characteristics of the NSTAR PPU and DCIU. *AIAA Paper No. 98-3938*, June 1998.
26. Christensen, J.A., et al. The NSTAR ion propulsion subsystem for DS1. *AIAA Paper No. 99-2972*, June 1999.
27. Ehricke, K.A. *Space flight*, Vol. II, *Dynamics*. Van Nostrand, Princeton, NJ, 1962, pp. 144–150.
28. Frick, R.H. and T.B. Garbon. Perturbations of a synchronous satellite. RAND Report No.R-399-NASA, May 1962.
29. Molitor, J.H., and M.H. Kaplan. Optimization of ion engine control systems for synchronous satellites. *AIAA Paper No. 63-273*, June 1963.
30. US Pat. 5,947,421, September 9, 1999, J.R. Beattie and P.J. Goswitz (to Hughes Electronics).
31. Molitor, J.H., et al. *J. Spacecraft Rockets* 4 (2): 176–182 (1967).
32. Seliger, R.L., and J.H. Molitor. Reliability in design: Solar-electric propulsion systems. *Ann. of Reliability Maintainability* 5: 256–273 (1966).
33. Molitor, J.H. Ion rocket systems and applications. *Modern Developments in Propulsion*, UCLA Course X458.1, July 1966.
34. Molitor, J.H., and R.N. Olson. Solar-electric propulsion systems for unmanned space missions. *AIAA Paper No. 67-713*, September 1967.
35. Spitzer, A. Near optimal transfer orbit trajectory using electric propulsion. *AAS Paper No. 95-215*, February 1995.
36. Molitor, J.H. Ion propulsion-technology spin offs. *AIAA Paper No. 76-1013*, November 1976.

JEROME H. MOLITOR
J.R. BEATTIE
Westlake Village,
California

RUSSIAN SPACE STATIONS

The Soviet and Russian long duration orbital space stations, Salyut and Mir, were operated from 1971 to 2001. During this period, eight stations were placed into orbit (Table 1). While the stations were in operation, a variety of improvements were made to extend their operational lifetimes and improve their performance. Experience in on-orbit operations enabled the Energia Rocket and Space Corporation (Energia RSC) to develop a Russian space-station concept. This concept was most fully embodied by the Mir space station and the Russian segment of the International Space Station.

The basic principles behind the Energia RSC concept, as actually developed and implemented, were as follows:

1. Use of a modular structure in which modules are classified as either “service modules” or “research modules.”
2. Use of a relatively inexpensive transportation infrastructure, including Soyuz manned spacecraft, Progress automated supply spacecraft, Soyuz automated spacecraft, and Raduga research-material return capsules, for station servicing.
3. All space-station equipment is designed for maintainability to increase the usable life span of the equipment.
4. All routine operation of onboard equipment is automated as much as possible to free the crew for research work.
5. For crew safety, a piloted spacecraft is always docked at the station.
6. The stations were used to address a broad range of research goals in science, engineering, and various applied fields.

The major goals of space-station use are as follows:

- research on the behavior of the human body under space conditions to determine a strategy for human space exploration (habitable structures in low Earth orbit, interplanetary flight, and lunar/planetary colonization);
- space research using instruments carried above the atmosphere;
- remote sensing of Earth’s surface and atmosphere for basic research as well as for applications, for example, studies of Earth’s natural resources for human use;
- basic research in materials science and the production of materials and biological substances under microgravity conditions;
- research and development of proposed materials and techniques for future projects (research on the behavior of complex structures; development of tugs, mechanisms, cable systems, and interplanetary spacecraft components).

The long duration orbital space station program began with the launch of the world’s first space station, Salyut, on 19 April 1971. At that time, it was still

Table 1. Operation of Soviet and Russian Orbital Space Stations

Space station	Date of flight	Date of active use	Number of primary crews	Number of visiting crews	Longest crew stay, days	Total duration of piloted flight, days	Prime developer
Salyut	19 Apr 1971–11 Oct 1971	19 Apr 1971–30 Jun 1971	1	—	22	22	Energia RSC
Salyut-2	03 Apr 1973–14 Apr 1973	—	—	—	—	—	NPOM
Salyut-3	25 Jun 1974–24 Jan 1975	25 Jun 1974–23 Sep 1975	1	—	14	14	NPOM-space station; Energia RSC-spacecraft
Salyut-4	26 Dec 1974–03 Feb 1977	26 Dec 1974–26 Jul 1975	2	—	63	91	Energia RSC
Salyut-5	22 Jun 1976–08 Aug 1977	22 Jun 1976–25 Feb 1977	2	—	48	65	NPOM-space station; Energia RSC-spacecraft
Salyut-6	29 Sep 1977–29 Jul 1982	29 Sep 1977–26 May 1981	5	11	184	668	Energia RSC
Salyut-7	19 Apr 1982–07 Feb 1991	19 Apr 1982–25 Jun 1986	5	5	237	873	Energia RSC
Mir	20 Feb 1986–23 Mar 2001	20 Feb 1986–23 Mar 2001	28	16	437	4591	Energia RSC-space station, spacecraft; Khrunichev modules

too early even to mention the idea of an operating concept for an orbital space station. Too many space station operational issues had not yet been resolved.

The issue of crew work assignments and capabilities during station operations had not been fully clarified. Another important consideration was that in 1971, it was still extremely difficult to predict either the station lifetime or the amount of time that crews could work on board the station.

The sole long duration flight before this was the Soyuz-9 spacecraft (crewed by A. Nikolaev and V. Sevast'yanov), which remained in orbit for 18 days; the scheduled duration of the Salyut mission was 3 months because at that time, it was virtually impossible to predict that a complex spacecraft could remain functional for several times that period.

The design for this first space station maximized the use of the design and equipment advances that were available at that time. Many of the on-board systems for the space station were adapted from the Soyuz spacecraft, and the pressure hull and basic structural elements of the main working compartment were adapted from the Almaz space station, which was then under development.

The orbital portion of the space station consisted of a transfer compartment, a working compartment, and a scientific equipment compartment. The 2.1-m-diameter working compartment had a docking system and a set of solar arrays, adapted from the Soyuz spacecraft. The working compartment, which was the largest, had a pressure shell that consisted of two hemispheres 2.9 and 4.1 m in diameter connected by a conical transition section. The 2.1-m-diameter equipment compartment was designed to house the orbital adjust propulsion adapted from the Soyuz spacecraft and a set of newly developed bipropellant attitude thrusters. A second set of solar panels (also adapted from the Soyuz spacecraft) was mounted on the exterior of the compartment. The scientific equipment bay was designed specifically for the Salyut space station, and was installed in the working compartment.

The attitude and flight control system (gyro units, integrator, computer, angular-velocity sensors, manual attitude control system, infrared vertical sensor, ion flux sensors, and pilot's scope), electrical power system (solar panels, standby chemical battery, charge control system), radio communications and telemetry system, radio orbit monitoring system, command uplink, central pilot control panel, rendezvous system, and life support systems were all adapted from the Soyuz spacecraft. Although the onboard control system was partially adapted from the Soyuz spacecraft, because of the large number of new systems and scientific instrumentation, in particular, it was substantially modified, and additional new equipment was added.

A new temperature control system was developed using the same hardware as that in the Soyuz spacecraft. One of the innovations involved the use of pipes carrying the heat-transfer agent to stabilize the hull temperature; this ensured favorable temperature conditions for the numerous hull gaskets, thereby maintaining integrity of the seals during a long flight.

The scientific equipment aboard the 1.3-metric-ton station included the following: a solar telescope, an X-ray telescope, an infrared telescope/spectrometer, and a 60× viewer, in addition to various other pieces of equipment. A significant fraction of the equipment was intended for astrophysical research. At

that time, it was believed that astrophysical research was the best route to new fundamental scientific results.

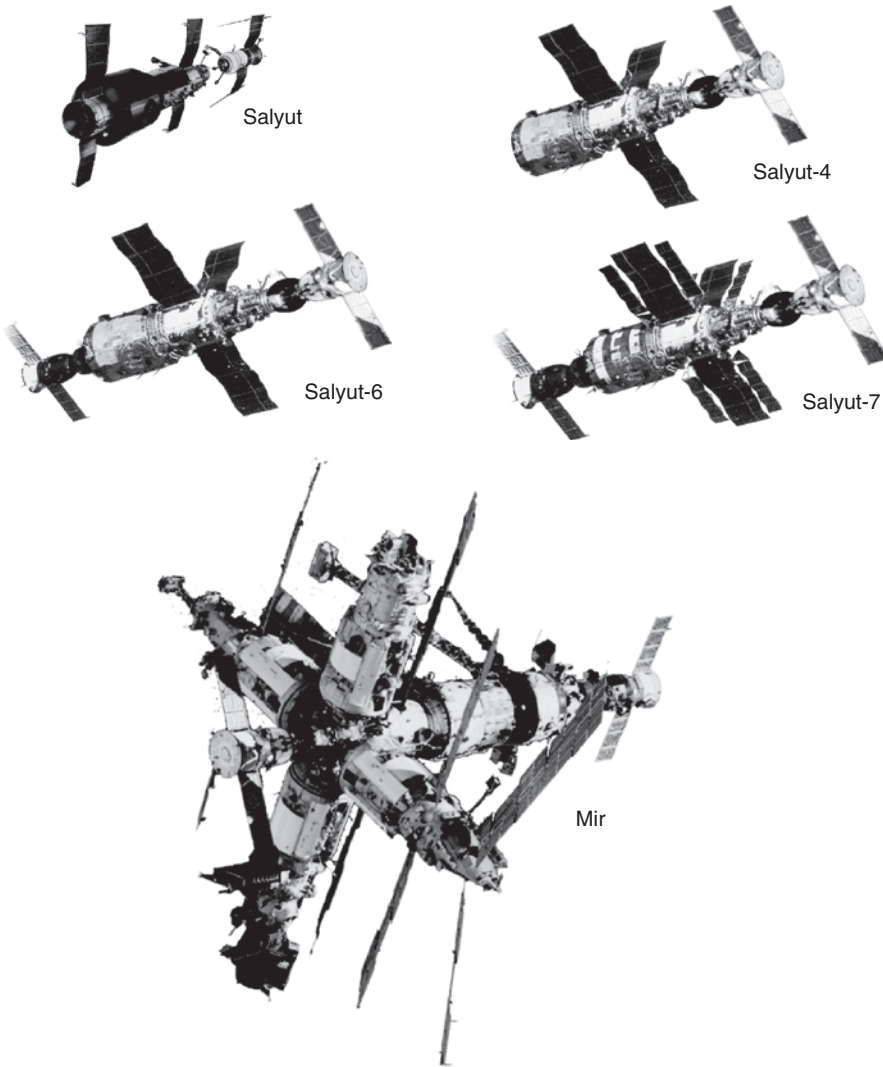
The Salyut space station hosted one expedition. A crew consisting of V. Dobrovolskii, V. Volkov, and V. Patsaev worked for 22 days in orbit. During this expedition, approximately 50 medical experiments were performed, including several experiments involving the cardiovascular system, collection of blood samples, and the blood supply to cerebral vessels, bone density and metabolic studies were also performed. The biological experiments were aimed at determining the effects of weightlessness on living organisms. One such experiment was the Oasis experiment. Seeds and plants were placed in a special container/greenhouse and were then studied and compared with experimental samples grown on Earth.

An extensive program of astrophysical research was scheduled; however, because the cover on the compartment containing the solar telescope and spectroscopic instrumentation did not open in orbit, most of the astrophysical research was canceled. The Orion telescope was used to map the sky at ultraviolet wavelengths that are invisible from the ground. Ultraviolet spectra of Vega and Hadar were obtained. Particle fluxes were determined using the Anna-3 gamma-ray telescope. Fixed and manually operated cameras were used to photograph Earth's surface in regions of greatest interest to geologists and cartographers, i.e. the mining region of Altai, Lake Balkhash, and central Asia. More than 1000 photographs of these regions were obtained.

Unfortunately, the first expedition to the first orbiting space station ended in catastrophe. The crew perished upon landing due to a failure in the Soyuz spacecraft. Significant improvements were made to the Soyuz spacecraft as a result of the accident, and work continued on the orbiting space stations (see Fig. 1). The first Salyut station was followed by the launch of the Salyut-2 and Salyut-3 space stations. These space stations were involved in specialized operations for the Ministry of Defense and were not part of the overall space station program.

The next space station in the program was Salyut-4, which was a significant improvement over the original Salyut space station. The main deficiency of the original Salyut space station was that it was necessary to expend a significant amount of fuel to keep the spacecraft rotating about an axis pointing toward the Sun during the main portion of the flight. This was necessary to keep the rigidly fixed solar arrays, which provided station power, pointing toward the Sun. Thus, the main change made in the space stations from Salyut-4 on (compared with the original Salyut space station) was the addition of three independently steerable solar arrays installed on the working module. To compensate for the increased mass of these solar arrays, the number of propulsion system tanks was reduced, and the station was moved to an altitude of 350 km to reduce fuel requirements for orbital maintenance. The mass of onboard research equipment was increased to 2 metric tons. The independently steerable solar arrays led to an improvement in research capability for Earth and space observations.

This station saw the first installation of the Kaskad cost-effective attitude control system and the Delta experimental navigational system. The temperature control system made the first use of an experimental heat pipe loop system, which turned out to be exceptionally promising for future generations of orbital



Specifications	Salyut	Salyut-4	Salyut-6	Salyut-7	Mir
Station mass, metric tons	25	25	36	36	140
Mass of research equipment, metric tons	1.3	2	2.4	2.5	11.5
Pressurized volume, m ³	92	92	98	98	440
Maximum power consumption, kW	3.5	4.1	4.1	5	45
Number of docking ports	1	2	2	2	3
Crew size, persons	3	3	3–6	3–6	3–6
Flight duration, yr	0.5	2.2	5	8.8	15
Service Spacecraft	Soyuz	Soyuz	Soyuz Progress	Soyuz Progress	Soyuz Progress Raduga

Figure 1. Russian and Soviet orbital space stations. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

space stations, and work was begun on developing of a closed-cycle water supply system for the crew, for which a condensate water regeneration system (CWRS) was installed on board the station. Approximately half the water excreted by the crew ends up in the air via the skin; the CWRS collects this water and treats it to make it potable again.

The Salyut-4 mission placed heavy emphasis on remote sensing of Earth's surface, and the mid- and southern-latitude territory of the Soviet Union was photographed. During these flights, valuable scientific data were obtained concerning the physical processes in the active region of the Sun, in Earth's atmosphere, and in outer space across a large portion of the electromagnetic spectrum. This marked the first time in the history of spaceflight that photographic and spectrographic observations were made of the auroras and noctilucent clouds (a rare natural phenomenon of great scientific interest). The effects of long duration spaceflight on the human body were studied, and various methods for treating the unfavorable effects of weightlessness were tested.

Engineering experiments to develop new systems and instruments for future spacecraft and long-term orbiting space stations formed an independent part of the flight plan. Successful performance of these experiments laid the foundation for further improvements in space hardware to be used for ever more complex tasks in space research. The assured on-orbit service life of the space station was extended from 90 days (the original Salyut space station) to half a year. However, the station actually remained functional for more than 2 years. Each successive space station was improved in the following areas: increased operational lifetime, improved capabilities for installation of special-purpose equipment, improvement of station performance for research (improved attitude stabilization, increased electrical power capacity, and improved control system).

Maintaining long duration functionality of an orbiting space station requires an economical supply mechanism. Development of the Progress automated cargo spacecraft was the principal design decision affecting the operational characteristics of subsequent generations of space stations.

The first Progress spacecraft docked with the Salyut-6 space station on 20 January 1978. It delivered consumables for the life-support system, additional research instrumentation, and additional fuel for the space station onboard propulsion system. At this point, it became possible to support effective multiyear station operations. The Salyut-6 space station included an additional transfer compartment with a second docking assembly enabling both a Soyuz spacecraft and a Progress cargo spacecraft to dock simultaneously at the station during resupply and cargo operations or two Soyuz spacecraft to dock at the station during crew changes. The second transfer compartment also provided an opportunity for extravehicular activities (EVA), in which case the transfer compartment became an airlock holding spacesuits and all equipment required to reach the station exterior; special handles on the station hull were available for the cosmonauts to move about and to attach themselves.

Various onboard systems were improved, a color television system was added, and a folding shower was installed in the operations compartment for enhanced crew comfort. The operational service life of the station was increased to 3 years. At this point, the space stations began to have a fairly mobile servicing

system. Soyuz spacecraft were used for station crew changes. This spacecraft was continually docked to the station and also served as an assured crew return vehicle. The assured crew return vehicle was simultaneously replaced as each crew changed.

All necessary cargo was transported to orbit using the Progress cargo spacecraft. The cargo manifest depends on the program and might include fuel, consumables for the life-support system, interchangeable equipment for scientific instruments, and/or interchangeable equipment for EVA.

An extremely important avenue for improving space station operating efficiency was opened once onboard maintenance and repair by the cosmonauts themselves became possible. This required overcoming the technological and psychological barriers related to crew safety during repairs while in orbit. The Salyut-6 scientific program resulted in the performance of more than 1550 experiments involving the use of more than 150 scientific instruments that weighed in excess of 2200 kg. More than, 750 kg of scientific instruments were brought on board using the cargo delivery capability. Research was conducted in the fields of astrophysics (submillimeter telescope and radio telescope with 10-meter parabolic dish antenna and others), materials production (Splav and Kristall experimental production systems), geophysics (photographic mapping equipment and multispectral camera), biology, medicine, etc.

One important mission of the orbiting space stations is studying the adaptation of the human body during long flights under weightless conditions. Such research is required to develop a strategy for future human participation in space exploration and can be performed only on board a space station. The longest stay on the Salyut-6 space station was 184 days.

The next space station in this generation was Salyut-7. In many respects, this station was identical to Salyut-6, but Salyut-7 also represented several engineering improvements. For improved efficiency of station operations, a special module contained scientific instrumentation (X-ray system, ultraviolet telescope, etc.), as well as a modern control system based on the Salyut-5B computer, and gyroscopes (flywheels) to enable attitude stabilization without fuel consumption. However, it was then decided to use this module on the Mir spacecraft instead. The module came to be known as the Kvant module.

During development of the Salyut-7 space station, it became clear that the solar array area needed to be increased. This change in station design did not seem possible due to problems in fitting the solar arrays under the nose fairing. It was then decided to retrofit the station with solar arrays while in orbit. A system of specialized cables for this purpose was included in the main solar array design, and these cables were used to install the additional panels, once the station was in orbit. Several serious repairs were made during the Salyut-7 flight. One of the fuel lines on the station lost pressure, perhaps due to a meteoroid impact. During a series of six spacewalks, the cosmonauts succeeded in identifying the failed line and hooked up a new line to the propulsion system. A special press for sealed line crimping and special valves mounted on the fuel nozzles, were developed.

In February 1985, during an unmanned portion of the flight, there was a loss of communications with the station due to a failure in the command system and an error by a Flight Control Center operator; the resulting inability to intervene in automated operations from the ground caused a failure in the

storage-battery charging mode, the system lost power, and the station failed completely. There was a real risk that the station might be completely lost. The main question was whether it would be possible to dock with a completely uncontrolled station so that repairs could be made. The future crew was given appropriate training to support docking with the station as an “uncooperative object.” No operation of this type had ever been performed before.

A special mission—cosmonauts V.A. Dzhanibekov and V.P. Savinyi on the Soyuz T-13 spacecraft—was sent into orbit. The crew rendezvoused and docked with the space station under manual control, using instructions from the ground, a laser rangefinder, and the onboard digital computer. The cosmonauts made the repairs, and the station was once again operational.

The main drawback of the Salyut stations was that they consisted only of a single module, which limited the options for placing research instrumentation. Another deficiency was a lack of electromechanical actuators, which implied high fuel consumption for attitude control and, thus, a large amount of cargo traffic from the ground to the space station. The Salyut space stations did not have any communications via relay satellites, meaning that there were large time intervals when the stations were out of communications range. The experience gained on the Salyut space stations with respect to control systems based on digital technology facilitated the transition to modern control systems.

The modular design of the new Mir space station led to enhanced options for housing a large quantity of scientific equipment. The space station consisted of six modules, the base unit and five research modules (Kvant, Kvant-2, Kristall, Spektr, and Priroda). The main service module in the station was the base unit and all of its support equipment. This module was based on the Salyut-7 base unit but had several significant improvements.

The following base-unit systems were modernized. The control system enhanced the capabilities of the station. The new Kurs rendezvous system did not require any rotation of the station during rendezvous operations. The power system capacity was increased, and voltage stabilization was improved. A radio system with a highly directional antenna was added for communications via a relay satellite.

The research modules were based on the FGB (functional cargo block) spacecraft developed by the Khrunichev State Center for Space Science and Space-Related Manufacturing. Initial plans called for attaching the Mir research modules to the station during the first year of flight. However, development and fabrication of the research modules turned out to be more labor-intensive than originally thought, and it actually took several years to install them. Despite this fact, the station continued in use, and the research program was designed around the modules that were present at the station at any given time.

The first research module at the Mir space station was the Kvant module, which included a complete astrophysical observatory with a system of telescopes developed in Russia as well as in various European countries. Note that the division into special-purpose modules and service modules is arbitrary because all modules contain both support equipment and research-oriented equipment. For example, the Kvant module included some support equipment that was vital to the space station as a whole: control-system units and flywheels for attitude stabilization based on gyroscopic forces to reduce fuel consumption. The awkward

atmospheric regenerators were replaced by a water electrolysis system to supply the crew with oxygen, as well as a regenerative carbon dioxide absorption system.

The next Mir space station module was the Kvant-2. This module contained research equipment to enhance the space station's scientific program. In particular, a steerable gimbaled platform carrying photometric, video, and spectrometric instrumentation was mounted on the exterior of the module. This platform could be controlled either by the space station crew or by Flight Control Center personnel. Scientists could now study, from the ground, any area on Earth's surface or in the sky. It included a fairly spacious airlock with a large, 1-m diameter exit hatch. The module included an additional set of flywheels installed on the exterior rather than in the interior (as in the Kvant module) of the habitable compartment. Unfortunately, further operational experience with the space station proved that this engineering design solution could not be justified because it turned out to be too difficult to replace the flywheels in the event of failure. This module also provided a new piece of propulsion equipment for use during EVA—the cosmonaut propulsion system (CPS), a “space motorcycle.” Several cosmonauts used the CPS, flew around the station, and photographed it.

The next research module in the Mir space station was the Kristall module. This module housed a wide variety of research equipment for use in materials-science research, including equipment for research on industrial-scale materials production under microgravity conditions and various pieces of biotechnology equipment. The Kristall module was equipped with a second docking assembly, the androgynous peripheral docking system (APDS) for docking with the Buran space shuttle.

Subsequent research modules, the Spektr and Priroda modules, were initially expected to be outfitted with spectral and laser equipment for remote-sensing observations of Earth, including observations aimed at studying Earth's natural resources.

The Russian-American Shuttle-Mir program, which occurred against the background of preparatory joint operations for the International Space Station, started before the launch of these modules. Research equipment to support work to be performed by an American astronaut was installed on the Spektr and Priroda modules. Once these modules were docked to the station, Mir began operations in its full configuration. The Shuttle-Mir program consisted of having American astronauts work on board the station; these astronauts were initially delivered via Soyuz spacecraft and eventually via the American Shuttle spacecraft. The American Shuttle spacecraft first docked with the Russian Mir space station on 1 July 1995. The Shuttle initially used the docking port on the Kristall module, but a special docking module with a special hatch was later installed at the space station. During a 3-year period beginning in 1995, the Shuttle docked with the Mir space station nine times as part of this Russian-American program.

June 1997 marked the most hazardous and troublesome event in the history of the Mir space station. During final experimental testing of a remote-control docking mode (manual control by a remote operator) with a Progress M-34 spacecraft, the crew was unable to reduce the approach velocity in time, the spacecraft crashed into the space station, and the Spektr module was depressurized. This module was isolated from the remainder of the station, thereby preserving the functionality of the station as a whole.

By 23 March 2000, the Mir space station had ceased active operations; in 2000, it was decided that it was not feasible to improve Mir operations at the same time that work was proceeding on the International Space Station. The Mir space station was deorbited and reentered the Earth's lower atmosphere over the South Pacific. During this 15-year period, the Mir space station performed an extensive program of research and experiments in various areas of science, engineering, and various areas of human activity.

The Kvant module X-ray telescope was the first to detect X-ray emission from Supernova 1987A in the Large Magellanic Cloud and follow the evolution of the supernova spectrum as a function of time. These observations had a high priority. An X-ray source in Cygnus (black-hole candidate) and clusters of galaxies in Perseus were observed. There were, in total, 6200 astrophysical observing sessions. A series of photographic cameras with different focal lengths were used to photograph Earth's surface at spatial resolutions of up to 10 m on multispectral film that had a maximum coverage width of 200 km on the ground. More than 125 million square kilometers of Russia and various foreign countries were photographed, resulting in more than 5000 photographic frames.

Several different control modes, including remote control from the Flight Control Center via a relay satellite, were tested on the spectroscopic instrumentation package that was mounted on the Kvant-2 steerable, gimballed platform. This remote-controlled, onboard detector system enabled the acquisition of real-time video and spectrometric data refreshed at 2-day intervals.

Plasma-beam ionospheric sounding experiments were performed, in part, to monitor electromagnetic-field variations as earthquake precursors.

While Mir was in operation, final modifications were made to various processes for producing epitaxial silicon structures, in addition to cadmium telluride, zinc oxide, and gallium arsenide monocrystals. Experiments were conducted for zone melting of germanium, silicon, and indium antimonide without crucibles. The semiconductor materials produced on board the Mir space station are being used for research studies, as well as for fabricating experimental devices and microelectronic components. There were more than 290 experiments involving manufacturing processes. Approximately 200 materials (both materials already in use and future materials) were tested during the space-based materials science experiments; the results of these experiments confirmed that space-related factors had a strong effect on structural materials and coatings. During these experiments, coatings for temperature control system radiators were studied, and it was confirmed that they would remain functional for up to 15 years; techniques and coatings were developed to protect materials against exposure to atomic oxygen by depositing thin films based on fluoroplastics, metal oxides, and silicon. In all, there were more than 2450 experimental sessions involving materials science.

Biotechnology is one promising field of research, and several experiments in this field were performed on various pieces of equipment aboard Mir. The most important of these were experiments on protein crystallization and the generation of highly effective producer cells through hybridization and electrophoresis in space. The high productivity and efficiency of the biotechnology processes have now been confirmed, and the necessary engineering data have been obtained to

construct a full-scale installation. In all, there were more than 130 experimental sessions involving biotechnology.

Experiments were performed on board Mir to study the growth and development of higher order plants as a function of time and to develop technologies using various growth media. An increase in biological activity by a factor of 5 or 6 relative to control samples occurred in crops exposed to radiation under space-flight conditions. Research was performed to study changes in the vestibular apparatus of various living organisms, quail, crabs, newts, and snails, under spaceflight conditions. In addition to their practical applications, the results obtained are also of great interest for basic science. More than 200 biological experiments were performed.

Numerous medical experiments and studies led to determination of the basic laws governing human adaptation to long-term space flight. The results enable us to predict with high confidence that longer and longer manned space flights will be possible, including flights to distant planets; the results will also find application in general medical practice. Doctor/cosmonaut V. Polyakov flew on Mir for 437 days. A total of 1400 medical experiments were performed on board Mir.

The engineering experiments enabled more precise determination of the specifications for newly developed structures. One such result was development of a reusable single-leaf solar array based on a hinged-rod beam for on-orbit use; the solar array was installed on the Kristall engineering module. A process and design was developed for a thin-film centrifugally spun 25-m-diameter mirror that could serve as the prototype for the primary element of a space-based system for nighttime illumination of Earth's surface by reflected sunlight.

A design was developed for a promising system (called the "Topol") for repeatedly opening and closing the solar panels in a two-leaf solar array for a future power system. A 6.4-m-diameter parabolic dish antenna was deployed at the space station, and it was confirmed that it would be possible to develop large antennas up to 100 meters in diameter. The experience in research on assembly and installation procedures for use in space, the physical and chemical properties of materials (including memory alloys), and the dynamics of large space-frame structures gained during preparations for, and performance of, these experiments suggests that it will be possible to develop specialized power modules with a load-bearing frame, a heat dissipation system, reusable solar arrays, and solar gas-turbine power plants. More than 800 engineering experiments were performed on board Mir.

Like the Salyut-6 and Salyut-7 space stations, the Mir space station also hosted cosmonauts from various countries in the Americas, Europe, and Asia, as well as experiments using foreign equipment. The equipment and instrumentation used also was of scientific and applied interest to the Russians and led to several novel, and, in some cases, important scientific and engineering results, as well as results in the applied sciences. In addition to the international scientific research collaborations, programs were also performed on Mir on a commercial or fee-for-services basis.

In 1990, the first protein crystallization experiment was performed pursuant to the terms of a commercial contract with the U.S. firm, Payload Systems. The Mir space station hosted cosmonauts from Japan, Great Britain, Austria, Germany, France (twice), and the European Space Agency (ESA) (German

citizens) on a commercial basis. Flights involving U.S. astronauts were also performed on a commercial basis. More than 31,200 experimental sessions were conducted at the Mir space station in a 15-year period.

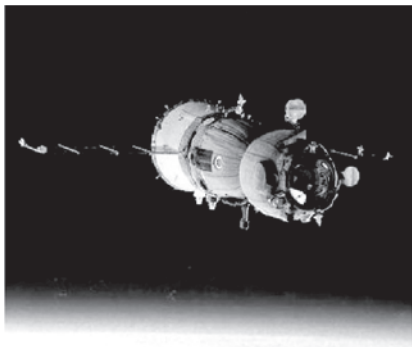
The Mir space station was a new-generation space station that absorbed all of the operational experience from orbiting space stations that had been developed in Russia over the years. The new engineering design solutions adopted for Mir enabled it to remain operational for a record period of time—15 years—an impressive achievement for an orbiting system of such complexity. Of course, it should be noted that the original developers of the space station did not anticipate that it would remain in operation for so long. Many interesting engineering design solutions were adopted for the space station and various related systems and assemblies. The system for supplying the space station with all required items—fuel and cargo—using Progress automated spacecraft is one of the essential features of Russian orbiting space stations. This system is unique in the world (see Fig. 2). Delivery of propellant for the space station thrusters was another logistical element mastered during space station operations. The two propellant components for the onboard propulsion system were transferred from tanks onboard the cargo ship using special airtight connectors.

Scientific research results from the Mir space station were transmitted to Earth via radio downlink or returned together with the crew on a Soyuz spacecraft. In addition, the space station also carried a special Raduga capsule that could be independently returned to Earth. This capsule was a small lander, thermally insulated for passage through the lower atmosphere at velocities nearly equal to the orbital velocity. The capsule is attached to a Progress cargo spacecraft and uses the Progress propulsion system for braking in Earth's atmosphere. After the spacecraft undocks from the station, it goes through a braking maneuver using its own propulsion system, whereupon the capsule separates from the spacecraft. The Progress spacecraft burns up in the atmosphere, and the capsule lands separately in the specified landing zone.

The solar arrays remained continuously oriented toward the Sun for the entire 15-year mission. The electrical drive for the solar arrays was located within the pressurized compartment, and the torque was transmitted to the solar-array structure through the airtight compartment wall by an electromagnetic field. This engineering design solution, first of all, enabled housing the electric motor under appropriate conditions, and, second, rendered a complex set of seals unnecessary for torque transfer to the solar-array structure on the space-station exterior.

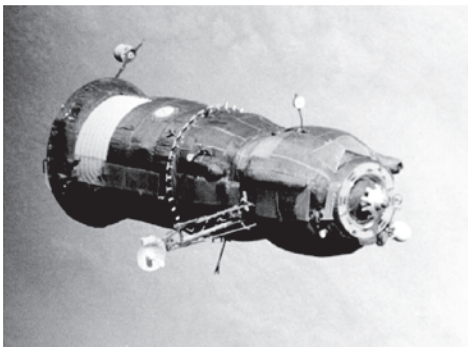
The Mir space station had a stand-alone propulsion system, which was mounted on the 15-meter Sofora truss. The propulsion system was mounted this far from the base module to increase the thrust lever arm for generating the required torques, and thereby reduced the amount of fuel required to eliminate accumulated angular momentum due to drag in Earth's upper atmosphere and gravitational perturbation torques. The space station control system had been improved from that used in the first Salyut. This system is used for space-station maneuvers and attitude control and determination of the space station's physical position (i.e., navigation). The Mir control system included a computer system and a set of sensors (Sun and star sensors and magnetometers). Gyrodynes

Manned Soyuz spacecraft



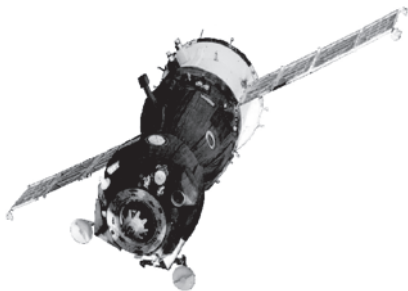
Spacecraft mass, kg	7100
Crew size	3
Delivery payload capacity, kg	150
Return payload capacity, kg	50
Maximum flight duration (on orbit), days	180

Progress cargo spacecraft



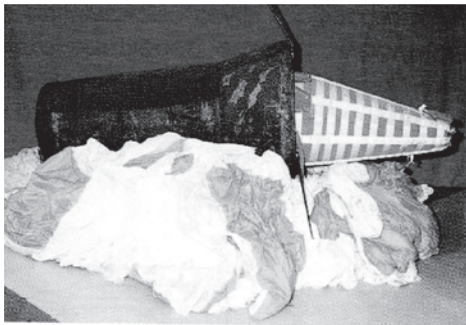
Spacecraft mass, kg	7300
Delivery payload and fuel capacity, kg	2500
Volume of cargo compartment, m ³	6
Fuel delivery capacity, kg	850
Maximum flight duration in orbit, days	180

Soyuz automated spacecraft



Spacecraft mass, kg	7100
Delivery payload capacity, kg	500
Return payload capacity, kg	500
Maximum flight duration (on orbit), days	180

Raduga return capsule



Capsule mass, kg	350
Return payload capacity, kg	150
Volume of return-cargo compartment, Liters	160

Figure 2. Transportation system used for servicing Mir space stations. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

(flywheels) were used as attitude-system effectors. Gyrodynes are single-degree-of-freedom power gyroscopes in magnetic bearings (for longer lifetimes and low noise levels). Standard procedure called for automatic docking with the Mir space station. However, the capability was always there to perform this operation

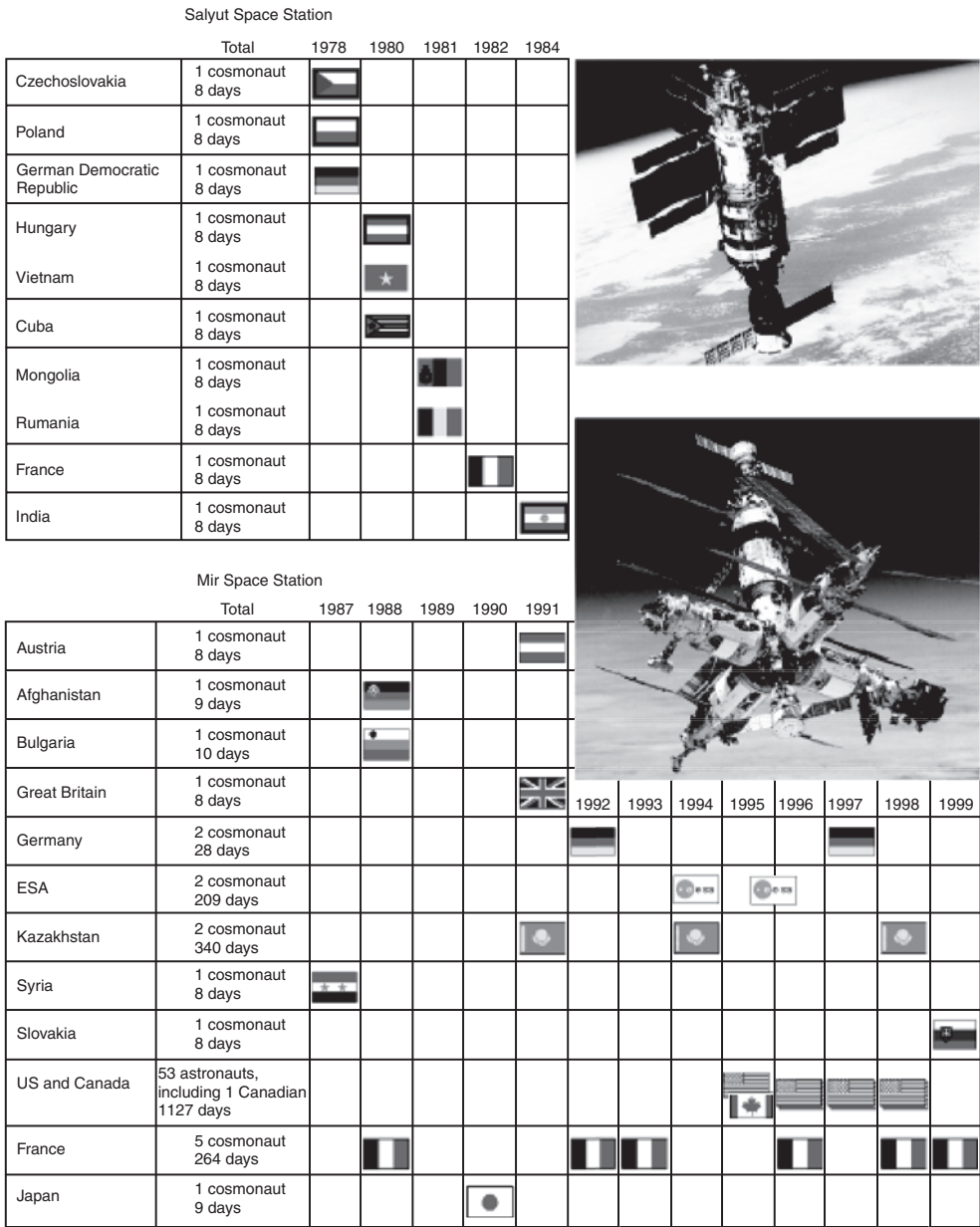


Figure 3. Salyut and Mir flights, 1978–1999. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

manually, and this capability was used on occasion. If necessary, a cosmonaut could dock the cargo spacecraft under remote control; the cosmonaut was on the space station, but his “eyes”—television cameras—were on the Progress spacecraft. A total of 75 extravehicular activities and three intravehicular activities (in the depressurized Spektr module), a total of 360 hours, were performed on the

Mir space station. The main purposes of these extravehicular/intravehicular activities were as follows: performance of experiments and research related to the behavior of materials in space, development of promising engineering design solutions, enlargement of the space station, and repairs.

The main deficiency of the Mir space station was limited power. And in spite of the fact that the space station had fairly large installed power in its full configuration, mutual shadowing of the solar panels prevented any fundamental solution of the power problem. The most effective technique was to mount the solar panels on masts extending outward from the basic structure and provide two degrees of freedom instead of one. Our operational experience with the Mir space station indicated that the Kvant-2, Kristall, and other research modules were too large, their weight, their cost, and the labor required to build them were not commensurate with their research equipment capacity. The modules should have used weight more effectively by using standardized bays for transporting modules to the space station.

One deficiency of the Mir space station was the low orbital inclination (51°), which put virtually the entire territory of Russia out of observational range. An orbital inclination of 68° would be required to increase the effectiveness of the space station in natural resources- and geophysics-related research. These changes were made in the design of the new Mir-2 space station. Additional changes include modifications to many of the onboard systems: a closed-loop crew oxygen and water supply system consisting of a water electrolysis system, a water condensate and urine regeneration system, and a system for recovering oxygen from carbon dioxide, and others. The control system was designed as an integrated data processing and control system. The power supply system used a combination of solar cells and a high-voltage (to minimize transmission losses) solar gas turbine with solar concentrator mirrors. Androgynous docking assemblies (APAS) were used.

The Mir-2 design became the design basis for the Russian segment of the International Space Station (ISS). The overall architecture, basic purposes, design of the modules, and equipment complement for the service systems in the Russian segment of the initial ISS configuration were similar to those used in Mir-2. The Salyut and Mir space stations became the first space laboratories to provide many countries around the world the opportunity to implement their own national space programs, because they were the first international space stations (see Fig. 3). The International Space Station project made full use of the operational experience gained on the Salyut and Mir space stations.

YURI P. SEMYONOV
L. GORSHKOV
ENERGIA RSC, Russia

RUSSIAN SPACEPORTS

Russia's spaceports (launch sites) serve as the basis for all Russian activities in space, enabling an independent policy in exploration and use of outer space. They comprise a number of facilities, equipment, and land areas designed for



Figure 1. The spaceports of the Russian Federation. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

receiving, storing, assembling, testing, prelaunch preparation, and launching rockets with spacecraft. Depending on its location, a spaceport may have one or several launch azimuths along which tracking stations are located.

The major facilities at spaceports are the launch and technical maintenance complexes, the fueling and decontamination stations, landing complexes, storage facilities of various types, launch debris impact areas, and telemetry monitoring stations. In addition, spaceports have plants for producing rocket propellants, airports, railroads, motor vehicle roads and various supply lines, as well as living quarters and social amenities.

Now Russia has three spaceports on its territory: Plesetsk, Kapustin Yar and Svobodnyy and, in addition, uses the Baikonur spaceport, which it acquired on a 20 year lease from the Republic of Kazakhstan in 1994. A map of these spaceport locations is provided in Fig. 1.

The History of Spaceport Construction in Russia

Spaceport construction followed logically from the need to ensure independent and efficient access to space for the Soviet Union and subsequently the Russian Federation. Spaceport construction started when a launch site was built for testing ballistic missiles in the region of the Kapustin Yar village in the Astrakhan Oblast. In 1946, a reconnaissance team headed by V.I. Voznyuk, selected the site of the future launch facility. The group studied seven regions of the USSR that seemed promising for this purpose and gathered and analyzed material related to the economy, weather, services, development level, construction potential, etc. Upon this group's recommendations the launch site was selected near the Kapustin Yar village on the lower Volga. V.I. Voznyuk was appointed the first head of the missile launch site. Under his leadership, the first launch of the R-1 ballistic missile took place from the Kapustin Yar launch site as early as

October 1947. Under the leadership of S.P. Korolev, between 1948 and 1956, the first Soviet ballistic and geophysical missiles were tested at this spaceport. Subsequently launch complexes were constructed for launching satellites for various purposes and also for vertical launch of the Vertikal geophysical missile. At the launch site an infrastructure was built to enable preparation work, launch spacecraft and provide all necessary services for the personnel to live at the site.

The injection of the first "small" artificial Earth satellite, later named Kosmos-1, on 16 March 1962 may be considered the point at which the Kapustin Yar launch site became a spaceport. Kapustin Yar became an international spaceport on 14 October 1969 when the first international satellite, Interkosmos-1 was launched.

The Kapustin Yar spaceport functioned as the launch facility for "small" missiles and "small" research satellites. This specialization continued until 1988, when the need to launch such satellites dropped sharply and launches from Kapustin Yar were terminated. However, the launch and technical complexes for Kosmos type launch vehicles were continuously maintained in working condition, which made it possible to use this site in 1996 to launch the German spacecraft ABRIXAS and MegSat.

In addition to construction of launch pads, the Znamensk town was built for spaceport personnel to live in. During different periods of time, V.I. Voznyuk, Yu.A. Pichugin, P.G. Degtyarenko, N.Ya Lopatin, N.V. Mazyarkin, V.K. Tonkikh, and V.P. Yushchenko headed the spaceport. The Kapustin Yar spaceport is currently under the authority of the Russian Ministry of Defense.

The next spaceport built was Baikonur. Construction of the Baikonur launch site was required to support the development of the R-7, the first Soviet intercontinental ballistic missile. Its flight range exceeded 8000 km and required equipping new launch azimuths eastwards through virtually the entire Asian part of the Soviet Union. New drop zones of LV separable parts and new telemetry stations were needed. A new launch site was required to launch these missiles. For this purpose, in the early 1950s, a commission was created to develop the specifications for the new launch site and recommend the most desirable areas of the country for constructing it.

The commission considered three locations for the launch site. As a result of their deliberations, a site in Kazakhstan in the vicinity of the Aral Sea and the town of Kzyl-Orda was selected. In February 1955, the USSR Council of Ministers adopted a Resolution concerning construction of the launch site near the Tyura-Tam railroad station in the Kzyl-Ordinsk Oblast of the Kazakh SSR. A.I. Nesterenko was appointed the first head of the site.

The major structures that had to be build were the launch pads for intercontinental ballistic missiles (ICBM) and the technical facilities of the launch site. By November 1956, construction of most of the facilities and installations were completed to provide flight tests of ICBMs. Facilities constructed included the launch complex, the launch control station, the assembly and test facility, some of telemetry stations, and the computer center building. The water line, water pumping station, and electric power station were all in operation. Motor vehicle roads and railroad lines were built.

On 15 May 1957, the first test launch of an R-7 ICMB took place but it failed. The second launch, planned for 9 June 1957 did not occur because

of defects in the missile equipment. The third missile launch took place on 12 June 1957 and was unsuccessful. The fourth and fifth launches occurred on 21 August, and 7 September 1957 and were finally successful. On 4 October 1957, the first artificial Earth satellite was launched from Baikonur and on 12 April 1961 a spacecraft carrying the planet's first cosmonaut, Yu.A. Gagarin was sent into space from there. Various launch and maintenance complexes were built at the launch site later. They were intended for preparing and launching spacecraft using launch vehicles of light (Tsiklon-M), intermediate (Soyuz, Molnia, Zenit), heavy (Proton) and super heavy classes, and the launch site infrastructure was developed. Now, all manned spacecraft and satellites into geostationary orbit (communications, television broadcast) are launched from Baikonur. This spaceport is also used for launching satellites into low and medium orbits (the GLONASS navigational system, Meteor weather satellite system, satellites for studying the Earth's natural resources, etc.), and also for launching unpiloted interplanetary probes and commercial spacecraft.

After the dissolution of the USSR, Baikonur spaceport became the property of the Republic of Kazakhstan and was leased by the Russian Federation. Now most of facilities at Baikonur are under the authority of the Russian Aviation and Space Agency (the launch complexes for the Soyuz, Molnia, Zenit, Tsiklon-M and Energia launch vehicles, one launch facility for the Proton, all assembly and test facilities, the oxygen-nitrogen plant). The Russian Ministry of Defense is responsible for the other Proton launch complex and the measurement complex. In 1990 the Baikonur Launch Site was renamed the Baikonur Spaceport.

The living area founded in 1955 was extended together with the spaceport. It had different names during its existence, and finally, in 1995, was renamed the Baikonur town. At different periods of time, the spaceport was headed by A.I. Nesterenko, K.V. Gerchik, A.G. Zakharov, A.A. Kurushin, V.I. Fadeyev, Yu.N. Sergunin, Yu.A. Zhukov, A.L. Kryshko, A.A. Shumilin, and L.T. Baranov. Now the Baikonur Federal Space Center, whose director is Ye.M. Kushnir, performs general work coordination of Russian Aviation and Space Agency facilities at Baikonur spaceport.

At different times throughout the spaceport's existence, the following leading scientists worked there: M.V. Keldysh, S.P. Korolev, I.V. Kurchatov, N.A. Pilyugin, M.K. Yangel, G.N. Babakin, V.N. Chelomey, V.P. Glushko, V.P. Barmin, V.N. Solovyev, B. N. Petrov, and V.F. Utkin.

For outstanding accomplishments in the testing of space technology, the following personnel of the spaceport were named Heroes of Socialist Labor: A.I. Nosov, A.S. Kirillov, V.A. Bokov, A.A. Shumilin, Ye.I. Nikolayev, and A.B. Berezin; The State Prize was awarded to B.A. Bululukov, A.P. Zavalishin, A.A. Shumilin, S.V. Limont, V.A. Menshikov, V.Ya. Khilchenko, Yu.N. Sergunin, V.A. Nedobezhkin, and V.I. Katayev.

Further development of space exploration and activities made it necessary to build the Plesetsk Spaceport in the northern regions of the USSR. At first, the Plesetsk Spaceport was established as a launch site at the Angara facility, where the first combat group of R-7 ICBMs was stationed. Construction of the Plesetsk launch site started in 1957. The launch site was built very rapidly in the severe northern climate. This work was headed by M.G. Grigoryev. During the 1960s

and 1970s, a highly developed infrastructure, including launch and technical complexes, was built to prepare and launch spacecraft using the Soyuz, Molnia, Kosmos, and Tsiklon launch vehicles.

The only missile launch site in Europe, Plesetsk enables to launch spacecraft for defense, socioeconomic, and scientific purposes and also in the framework of international collaboration programs. The Mirnyy town was founded simultaneously with construction of the launch site facilities and implementation of space programs. Plesetsk launch site was granted the status of Spaceport in November 1994. During various periods of time, the launch site was headed by M.G. Grigoryev, S.F. Shtanko, G.Ye. Alpaidze, Yu.A. Yashin, V.L. Ivanov, G.A. Kolesnikov, I.I. Oleynik, A.N. Perminov, A.F. Ovchinnikov, G.N. Kovalenko. Now, Plesetsk is under the authority of the Russian Ministry of Defense.

The Svobodnyy Spaceport was founded in 1993 and received official spaceport status in March 1996. The Svobodnyy Spaceport was formed using the facilities of a decommissioned missile division in the Amur Oblast. The existing infrastructure of the spaceport had to be modified for launching light-class launch vehicles derived from decommissioned ICBMs. The first head of the spaceport was A.A. Benediktov.

Current Russian Activity in Space

Space activity includes all activities associated with exploration and use of outer space, including the Moon and other celestial bodies. Space activity occupies one of the key places in Russian geopolitics and is the most important factor determining the country's status as a world power and a country possessing highly advanced technologies. The exploration and use of outer space plays an ever more important role in the economic, scientific, and social development of the country and in national security. The characteristics of the Russian Federation's geographic position (its size, the length of its sea, land and air borders, its varied terrain, rich natural resources, and other factors) have led logically to develop and efficiently use its space potential.

Russian space activity is implemented in accordance with the Russian Federation Law "About Space Activity." The major goals and objectives of RF space activity are defined by the Concept of the Russian Federation's National Space Policy. The main objectives of Russia's space policy are modernization, consolidation and efficient use of its space potential to increase the economic and defensive power of the country, ensure its national security, develop science and technology, solve social problems and expand international collaboration.

The directions of space activities are defined and developed in Russia's Federal Space Program approved by the RF Government every 5 years.

Space activity includes creating (including development, manufacture, and testing) and using space technology, materials and technological processes, and providing other services associated with space activity, and also the RF's international collaboration in exploration and use of space.

Volume of Freight Traffic from Russia's Spaceports Compared to Other Spaceports in the World. Between 1957 and 2000 throughout the world, there were approximately 4000 rocket launches for exploration and use of space.

More than 2500 were launched from the spaceports of Russia (Table 1). Plesetsk was the leader in the number of launch vehicles sent into space. However, Baikonur was the leader with respect to the volume of freight traffic. Up to 80% of the total payload launched annually from all of Russia's spaceports was launched from Baikonur. Between 1995 and 2000 Baikonur occupied first place in the number of launches as well. During this period, approximately 70% of the total number of Russian launches took place from this spaceport. The maximum possible payload traffic for Russian spaceports is approximately 60% of the total payload traffic from all launch sites of the world.

The Main Launch Vehicles Used at Russian Spaceports. The launch vehicles currently used at the Baikonur, Plesetsk, Kapustin Yar and Svobodnyy spaceports to launch spacecraft differ in classes and types and provide injection into near-Earth orbit of payloads weighing between 50 kg and 20 tons. The existing system of launch vehicles includes the expendable Kosmos, Tsiklon, Tsiklon-M, Start, Rokot, Dnepr (light class); Soyuz, Molnia, Zenit (intermediate class); Proton (heavy class); and also launch vehicles based on converted ICBMs.

The main types of launch vehicles used in Russian spaceports are cited in Table 2. Information about the number of launches of the major launch vehicles from Russian spaceports as of January 2000 is provided in Table 3.

In Russia currently, the basic launch vehicles are Soyuz and Proton types. These account for most of the spacecraft launches. Their high reliability and relatively low cost of manufacture compared to foreign analogs have made these launch vehicles competitive on the world market for launch services.

Space launch vehicle systems using the Soyuz intermediate-class launch vehicle are assembled at the Baikonur and Plesetsk spaceports. The Soyuz was developed from the R-7A ICBM. In the past, this launch vehicle was the foundation of intermediate-class Russian launchers for manned and unmanned spacecraft used for various purposes.

The intermediate-class Soyuz launcher uses ecologically clean propellants, kerosene and liquid oxygen. The liftoff mass of this vehicle is approximately 310 tons, and its engines have a total thrust at sea level up to 400 tons. Its technical specifications make it possible to insert payloads of up to 7 tons into a reference orbit. The Soyuz is one of the most reliable and efficient launch vehicles in the world. Its reliability has reached a value above 98%. The Soyuz is used to launch approximately 40% of all Russian spacecraft launched annually.

Three-stage heavy-class Proton launch vehicle (used only at Baikonur) can insert a payload of up to 20 tons into a reference orbit, and when the DM upper stage is combined with it, it can put a satellite weighing up to 3.5 tons into geostationary orbit. The thrust of its engines at sea level is 900 tons. Its liftoff mass is 690 tons, its length 44.3 meters, and its maximum cross section is 7.4 meters. The propellants used in this launch vehicle are unsymmetrical dimethyl hydrazine (fuel) (UDMH) and nitrogen tetroxide (oxidizer). At present, Proton launch vehicles are being updated by installing an improved control system and a new upper stage, Briz-M. As a result, environmental pollution with propellant traces in the launch debris impact area will be decreased, and the impact area will be significantly diminished. The updated Proton-M with the new upper stage, Briz-M, was successfully launched on 7 April 2001.

Table 1. Number of Launches from the Spaceports of the World and their Potential Payload Capacities

Country	Spaceport	Number of launches by year						Max. payload capacity to low orbit, tons/yr
		1957–1994	1995	1996	1997	1998	1999	2000
Russia	Baikonur Plesetsk Kapustin Yra Svobodnyy	968	19	16	18	17	21	30
		1413	13	11	9	7	6	5
		84	–	–	–	1	2	–
United States	Eastern Test Range (Cape Canaveral) Kennedy Space Center Western Test Range (Vandenberg AFB)	–	–	–	2	–	–	1
		494	16	17	16	17	16	18
		”	7	7	8	5	3	
China	Jiuquan Xichang Taiquan	503	4	4	9	8	11	10
		22	–	1	1	–	1	–
		15	3	3	3	2	2	4
Japan	Tanegashima Kagoshima	2	–	–	–	4	1	1
		26	1	1	1	1	–	–
France	Kourou	21	1	–	1	1	–	2
		69	11	11	12	11	10	12
India	Sriharikota							
		6	–	1	1	–	1	1
								4.5

Table 2. The Main Types of Launchers Used at Russian Spaceports

Launch vehicle class	Spaceport			
	Baikonur	Plesetsk	Kapustin Yra	Svobodnyy
Light class	Tsiklon-M Rokot Dnepr	Kosmos Tsiklon Start Rokot	Kosmos	Start
Intermediate class	Soyuz Zenit	Soyuz Molnia		
Heavy class	Proton			

The Zenith intermediate-class launch vehicle (launched at Baikonur) is used to insert spacecraft into low orbits, including sun-synchronous ones. It has two stages and is capable to insert a payload mass up to 13.7 tons into a reference orbit at an altitude of 200 km and inclination of 51°. The liftoff mass of the launch vehicle is 460 tons. Both stages use ecologically clean propellants—liquid oxygen (oxidizer) and kerosene (fuel). All operations to prepare the vehicle for launch are automated. If the launch is canceled, work to return the vehicle to its initial state is implemented by remote control from the command point. Modified Zenit stages were used as the side modules of the Energia launch vehicle. In the late 1980s, space programs were seriously curtailed. Many new satellites were not built. For this reason, there were only 32 launches of Zenit LV.

Light-class Tsiklon-M launch vehicles (Baikonur launches) based on converted R-36 ICBMs, were developed by the Yuzhnoye Design Bureau under the direction of M.K. Yangel, General Designer. The liftoff mass (not counting the spacecraft mass) is 178.6 tons. This launch vehicle enables inserting spacecraft weighing 3.2 tons and 2.7 tons into circular orbits of 200 km at inclination of 65° and 90°, respectively. Launches of Tsiklon-M began in 1967. At present, the launch vehicle is used only for launching Kosmos series spacecraft.

Table 3. Number of Launches as of 1 January 2000

Spaceport	Year of first launch	Number of launches	Number of failures
Heavy-class launch vehicles			
Baikonur	1965	266	11
Intermediate-class launch vehicles			
Baikonur Plesetsk	1957	705	56
	1966	904	32
Light-class launch vehicles			
Baikonur Plesetsk Kapustin Yra	1967	103	0
	1973	515	22
	1967	87	2

The three-stage light-class Tsiklon launch vehicle has been used at the Plesetsk spaceport since 1977. It has a lift-off mass of 191 tons. It is capable to insert a payload of up to 4 tons into high orbit of 200 km and launch six spacecraft at a time. The launcher uses toxic propellants: nitrogen tetroxide (oxidizer) and UDMH (fuel). When the Tsiklon launch vehicle system was developed, new approaches for preparing a vehicle for launch were incorporated. This advanced Soviet rocket building to a new level in the mid-1960s. The cycle of prelift-off preparation and launch of Tsiklon is 100% automated, and all operations at the complex is at least 80% automated. The developer and manufacturer of the Tsiklon vehicle is the Yuzhnoye NPO (Ukraine).

The two-stage light class Kosmos launch vehicle (Plesetsk) enables inserting payloads of up to 1500 kg into circular orbits from 200 to 2000 km in altitude and elliptical orbits. The vehicle liftoff mass is up to 109 tons. It is 32.4 meters long and 2.4 m in diameter. The engine units of the launch vehicle run on hypergolic propellants—an oxidizer (nitric acid) and UDMH fuel.

The Start launch vehicle was developed by the Kompleks-MIT Scientific and Technological Center to launch small spacecraft using mobile launchers. At present, this vehicle is used for launches of spacecraft of 200–400 kg from the Svobodnyy spaceport.

The Khrunichev State Research and Production Space Center developed the Rokot launch vehicle. The launch vehicle, which is based on the RS-18 military missile, uses the new Briz-KM upper stage that can employ complex orbital insertion patterns. This is especially important for limited selection of drop zones for LV separable parts.

The Dnepr launch vehicle was developed by the Kosmotras International Space Company on the basis of RS-20 military missiles to inject single and multiple payloads into space from Baikonur. The basic version differs minimally from the standard RS-20 missile. Fitting the vehicle with a more powerful third stage is being considered.

Baikonur Spaceport

Baikonur Spaceport is located within the Republic of Kazakhstan. The geographical coordinates of the spaceport are 46° North latitude and 63° East longitude. A schematic map of the spaceport is presented in Fig. 2. The total area of the spaceport is 6717 km². In accordance with a treaty between the Russian Federation and the Republic of Kazakhstan on the basic principles and conditions for use of Baikonur, signed by the heads of state on 28 March 1994, Baikonur was leased to the RF for 20 years. To ensure effective use of Baikonur spaceport to implement various space programs, Russia and Kazakhstan have signed appropriate agreements concerning lease conditions. This has made it possible to create a legal foundation for the Russian Federation's use of Baikonur, as well as the most favorable conditions for efficiently investing funds in the spaceport's ground-based infrastructure, which is necessary to implement space programs. The property rights to the real and movable property built, acquired, and delivered by the Russian Federation to Baikonur belong to it, as the party providing the funds.

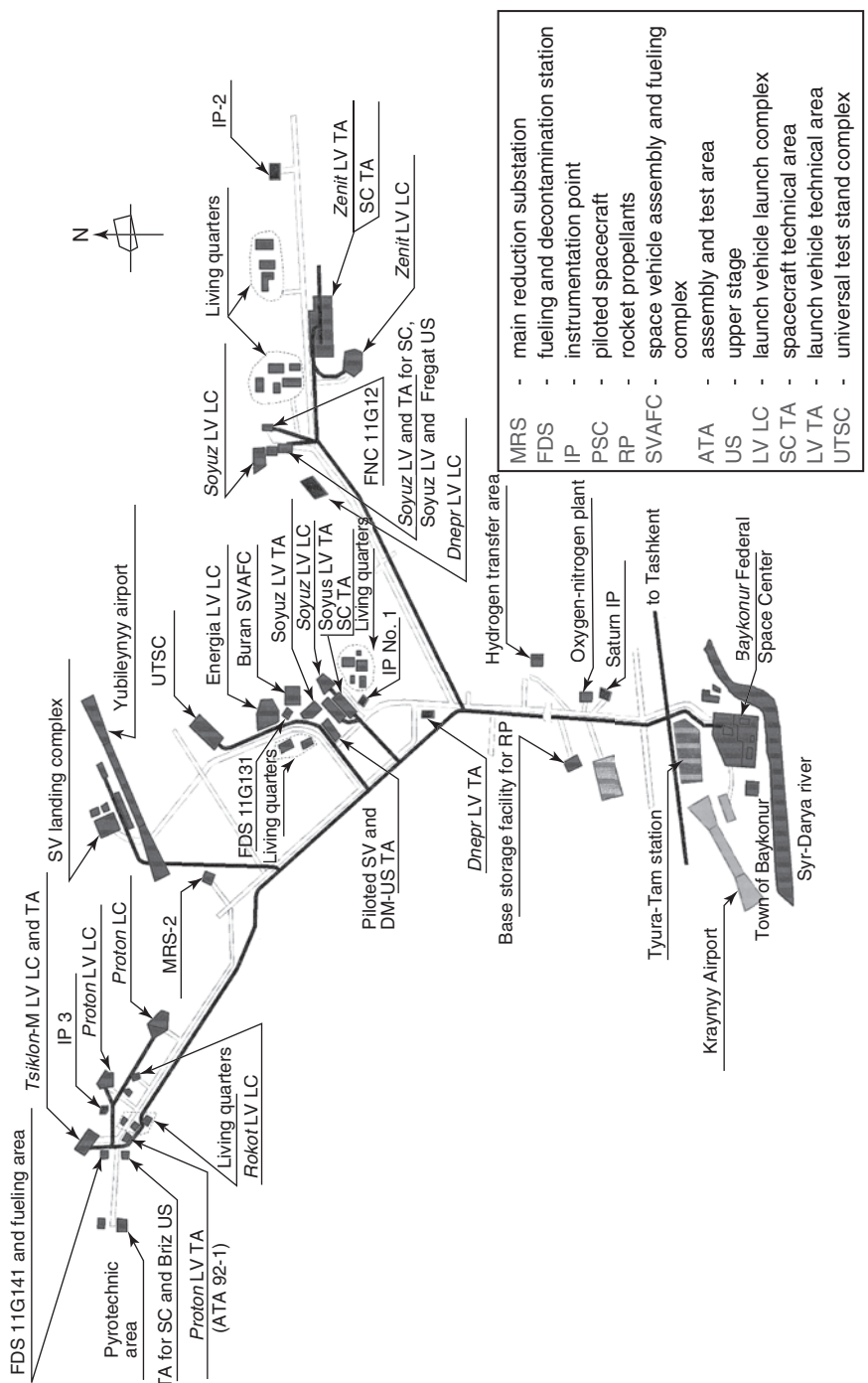


Figure 2. Schematic map of Baikonur. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Before 1993, Baikonur Spaceport was fully under the authority of the Russian Federation Ministry of Defense. Since 1993, the Russian Space Agency, subsequently the Russian Aviation and Space Agency, has been taking an active part to ensure functioning and use of all facilities. From 1993 to 2000, approximately 80% of the spaceport infrastructure, which is used by space industry organizations, was transferred to the Russian Aviation and Space Agency.

Baikonur spaceport includes nine launch complexes with 15 launch pads and 11 assembly and test areas with 34 technical complexes for assembly, testing and prelaunch preparation of launch vehicles and spacecraft; three fueling and decontamination stations for fueling spacecraft and upper stages with propellants and compressed gas; a measurement complex with a powerful computer center, and an oxygen and nitrogen plant with a total production capacity of up to 300 tons of cryogenic products per day. The spaceport infrastructure includes an advanced electric power supply grid, which contains more than 600 transformer substations and 6000 km of electric transmission lines; two first-class airports, including Yubileyny, can receive different classes of aircraft; more than 400 km of railroad tracks; more than 1000 km of motor roads and 2500 km of communications lines.

The Russian Federation uses Baikonur spaceport to satisfy the demands of the state for space-based means of communications; television and radio broadcasting; remote sensing of the Earth; time standards and geographic navigational coordinates of various types to users on the ground, in the air and at sea; and implementation of manned and international collaborative space programs.

The Baikonur spaceport plays an extremely important role in implementing Russia's space programs, since it occupies the leading place among Russian spaceports with regard to the number of space launches. Thus, from 1996–1999, 70% of all Russian launches took place from Baikonur; the analogous figure for 2000 was 83%. Spacecraft of various purposes and automated interplanetary probes are prepared and launched from Baikonur into orbits from 200 km to 40000 km high using light-intermediate- and heavy-class launch vehicles. The azimuth of launches has ranged from 35 to 192°.

To support preparation and implementation of spacecraft launches, Baikonur has space rocket complexes for the Proton, Soyuz, Zenit, Tsiklon-M, and Rokot vehicles. These complexes comprise launch and technical facilities for preparing launch vehicles, upper stages, and spacecraft. The technical facilities are located in the assembly and test areas with sites for assembly and testing of launch vehicles and spacecraft, fitted with all necessary technical, installation, rigging, docking, and crane equipment.

Baikonur is a unique spaceport due to launch facilities for heavy Proton launch vehicles, which are used to launch all space station modules, including those for the International Space Station (ISS), Russian television broadcasting satellites, most of the relay, communications, and navigation satellites, and also interplanetary probes. The launch facilities for Proton were developed under the direction of General Designer V.P. Barmin. The first launch of a Proton vehicle with a heavy spacecraft occurred on 16 July 1965. The weighty contribution of Baikonur to the conquest of space involved Proton launches with interplanetary probes to the Moon, Venus, Mars, and the Salyut and Mir orbital stations, as well as launches of Ekran and Gorizont satellites for national economy programs.

Both transport and cargo vehicles in the framework of manned flight programs were launched from Baikonur.

The space rocket complexes developed at Baikonur for Soyuz launch vehicles are well known throughout the world due to launching all manned spacecraft with Russian and international crews. The Design Bureau of General Machine Building developed these legendary launch complexes under the direction of General Designer V.P. Barmin.

The contribution of Baikonur to the implementation of Russia's national space programs may be evaluated by the number of spacecraft launched in each of these programs. Thus, there were launched from Baikonur: 95% of spacecraft in the Earth space observation and remote sensing programs; more than 70% of the spacecraft in the navigational program; 25% in the communications and television broadcasting programs; more than 40% in the weather satellite program; more than 30% of the spacecraft in the scientific research program; and 100% spacecraft in the manned spacecraft program. The Baikonur Spaceport was used to implement such space programs and missions as Vostok, Voskhod, Salyut, Mir, Mars, Venera, Luna, and Energia-Buran. Baikonur played the largest role in the implementation of the Mir program. Approximately 220 organizations and 80 scientific-research institutions participated in the Mir program.

The history of the spaceport is associated with the construction of launchers. The first launcher for the Soyuz launch vehicle was put into operation in 1957. The second, analogous to it, in 1961. Two launchers for Tsiklon-M were put into operation in 1967. The first Proton launcher went into operation in 1965, the second in 1966, and two more in 1979.

Launch complexes for the Energia launch vehicle (General Designer B.P. Barmin) were built at the spaceport in the framework of the Energia-Buran program. On 15 May 1987, these facilities provided the successful test launch the new superpowerful Energia launch vehicle from Baikonur. On 15 November 1988, this launch vehicle was used to insert the 30-ton reusable orbital spacecraft Buran (unmanned version) into near-Earth orbit. The success of this outstanding experiment was due to the ground-based test facilities for this program that already existed at this spaceport. The Energia-Buran program was preceded by the grandiose N-1 lunar program adopted by the USSR government in 1964. Between 1969 and 1972, there were four launches of the N-1 launch vehicle, which were unsuccessful. In 1976, work related to N-1 launch vehicle was terminated completely. The unique ground-based complex created for this program was closed down, but at the end of the 1970s, rebuilding was started and new ground-based facilities for the Energia-Buran complex began to develop. As a result, a launch complex and a universal multisystem launch stand were built. A special landing facility was built to allow the orbital spacecraft to land at Baikonur. In the early 1990s, the Energia-Buran program at the spaceport was terminated, the facilities were closed down and some of them were used for other programs.

At the same time as the Energia-Buran complex was being built at Baikonur, a ground complex was also built for Zenit—a new generation of middle class launch vehicles—which could insert a payload of 15 tons into near-Earth orbit. A key feature of the Zenit complex is the maximum level of prelaunch automated operations. Zenit's launch facility was developed by the Design Bureau for Transport Machine Building under the direction of Chief Designer V.N. Solovyev.

Silos for Rokot and Dnepr were reequipped to launch spacecraft. The first launch of Rokot occurred in 1994, and the first one of Dnepr—in 2000.

The Baikonur Spaceport is used for international programs in which Russia participates. These include: Phobos, Vega, Interkosmos, IRS, and the International Space Station. Plans for collaboration with the world community for developing and using the International Space Station were designed specifically for Baikonur. In recent years Baikonur began to be used extensively for commercial space projects. The Proton and Soyuz launch vehicles have the same capacities as their Western analogs, and in the near future, we can expect expansion of their use for spacecraft launches in the world market. The first two commercial launches occurred in 1996; in 1997-2000 there were already 33 spacecraft launched from Baikonur. Foreign partners actively collaborate with Russian enterprises in marketing these rockets in the world market.

The practical implementation of the principles of Russia's international collaboration in the launch services market started with establishing of the ILS joint venture based on the Lockheed-Khrunichev-Energia International, Inc. As a result, the Proton space rocket system entered the market and demonstrated its capabilities.

The Starsem Company in the Globalstar and Cluster programs used the facilities for the Soyuz launch vehicle. In the framework of these programs there were 10 launches of Soyuz LV since 1999.

The Baikonur Spaceport is a component of the Baikonur Complex, which also includes an administrative center—the Baikonur town. The infrastructure of the Baikonur town includes more than 300 apartment buildings, six hotels, a hospital with 1600 beds, an inpatient clinic with 360 beds, and two out-patient clinics that handle 470 and 480 patients a day. The town has a whole series of educational institutions: more than 10 schools for general education, a branch of the Moscow Aviation Institute, a communications technical school, a medical school, and others.

The future prospects for Baikonur rest on technical modifications of the launch and technical complexes for Proton LV to support launches of Proton-M, including launches involving the Briz-M, DM-03 and KVRB upper stages; updating of the launch and technical complexes for Soyuz to support launching of Soyuz-2, including launches with the Fregat upper stage; reequipping the infrastructure of the Zenit launch complex for launching Zenit with its new control system; and developing of new work sites for preparing and testing spacecraft in the framework of the Russian Federal Space Program.

Plesetsk Spaceport

The Plesetsk Spaceport is located in the Arkhangelsk Oblast and has geographical coordinates of 63° North latitude and 41° East longitude. A schematic map of the complex is provided in Fig. 3. The total area of the complex is 1762 km². The Plesetsk Spaceport is used for launching spacecraft for scientific, social and economic, and defense purposes and also in the framework of international and

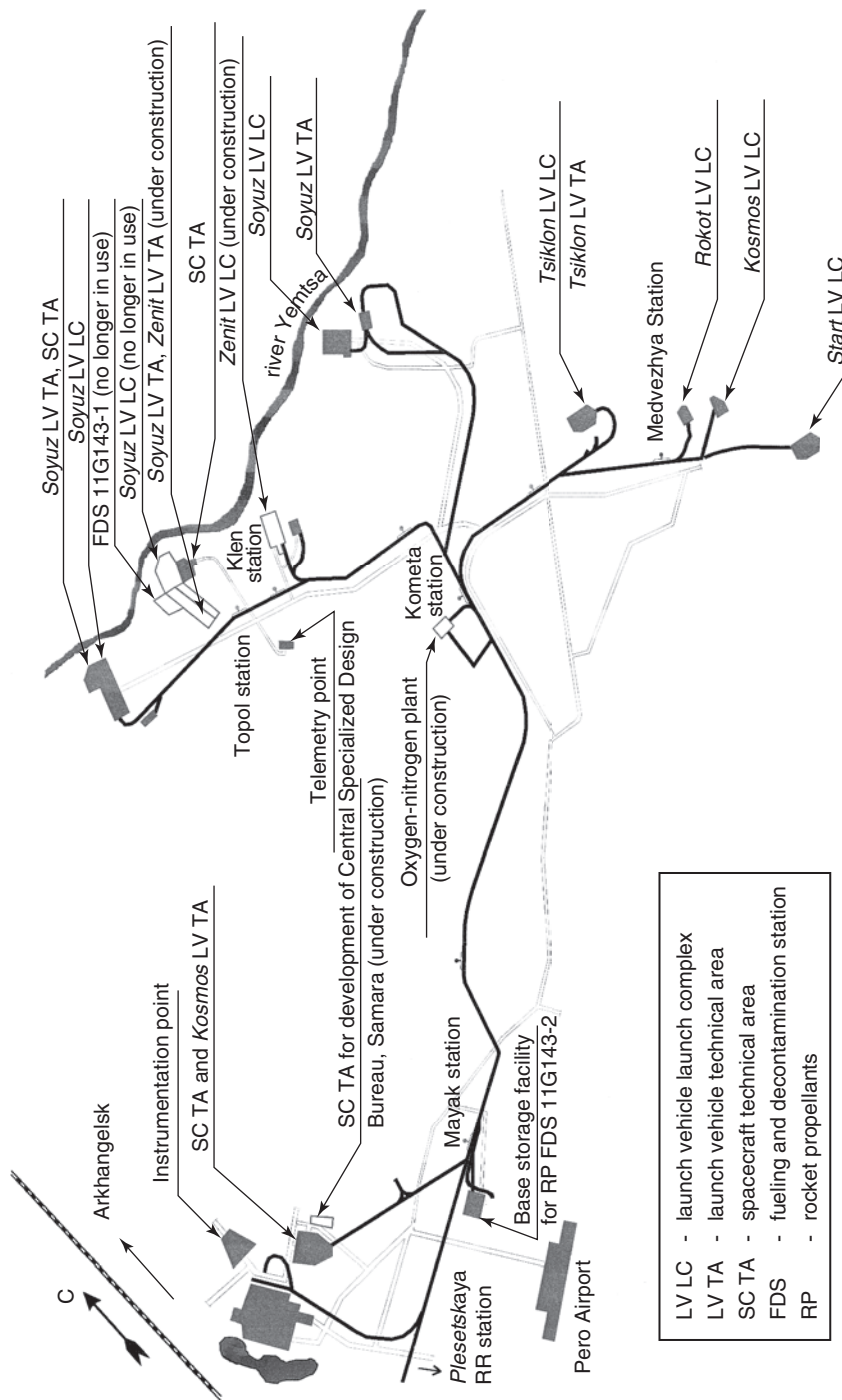


Figure 3. Schematic map of the Plesetsk spaceport. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

commercial space programs. Plesetsk has six launch complexes with nine launch pads, six assembly-test areas, containing 37 technical complexes for assembly, testing, and preparation of launch vehicles and spacecraft; two fueling and decontamination stations; and a measurement complex with a computer center. The spaceport's infrastructure includes a well-developed network of power lines, whose total line extent is 152.4 km, railroads and asphalt-concrete roads, a water and heat supply network, and the Pero airport. As the major site of space launch on Russian territory, the spaceport plays a key role in providing the country with independent and free access to space.

During Plesetsk's existence, approximately 1500 launches of space rockets took place from there. This represents approximately 60% of all space rocket launches from Russian spaceports. In the 1980s, up to 60% of all space rockets were launched from Plesetsk. In the 1990s, the number of launches of space rockets decreased, and between 1996 and 1999, this was approximately 30%.

The role of the Plesetsk spaceport in implementing Russia's national space programs can be evaluated by the number of launches for each program. Thus, Plesetsk launched 35% spacecraft for the space observation program, 75% for communications and television broadcasting, more than 25% spacecraft for navigation, approximately 70% spacecraft for the remote Earth sensing program, approximately 60% spacecraft for the weather satellite program, and up to 70% for the scientific research program.

Plesetsk's participation in the implementation of national space programs has been associated with such projects as Kosmos, Molnia, Tsikada, Meteor, Okean, Foton, AUOS, and Prognoz. The international missions launched from Plesetsk include: MAS-1, Magion, Gelisat, Tubsat, Astrid, Faysat, COSPAS-SARSAT, and Bion. The history of the Plesetsk launch site as an official spaceport began with the rebuilding of the R7-1 ICMB launch facilities for Soyuz and Molniya launch vehicles (General Designer, B.P. Barmin) and with founding in 1964 of a spacecraft testing and launch administration.

On 17 March 1966, the first spacecraft, called Kosmos-112 was launched from launch pad No. 1 of the Angara facility. As the number and type of satellite launched in the Soviet Union increased, the process to build new launch and technical complexes continued at Plesetsk spaceport. The first standard launch complexes specially for Kosmos launch vehicles were developed and put into operation in the last half of the 1960s and in the early 1970s (Chief Designer, V.N. Solovyev). More than 450 space rockets were launched from these complexes. Now one of these was refitted to launch Rokot launch vehicles. In the 1970s, launch and technical facilities for the Tsiklon-3 vehicle were built and put into operation in 1980 at Plesetsk, and since then they have been used for hundreds of launches.

The need to improve existing space rocket complexes and develop new ones determine the future prospects for Plesetsk, including updating the launch and technical complex of the Soyuz vehicle to prepare and launch Soyuz-2, and the refitting of the Kosmos launch facility for Rokot LV, supporting measures to launch Start LV, development of a unified launch and technical complex to prepare and launch Angara light- and heavy-class vehicles, and development of technical complexes for future spacecraft.

Svobodnyy Spaceport

This spaceport is located in the Svobodnensk Region of the Amur Oblast in Khabarovskiy Kray on a site whose geographical coordinates are 51° North latitude and 128° East longitude. A schematic map of the complex can be seen in Fig. 4. The total area of the spaceport is approximately 972 km². The spaceport infrastructure includes technical complexes for launch vehicles and spacecraft, the launch facility (launch pad and temporary command post), the instrumentation stations along the launch azimuth, the communications and data transmission lines, motor roads and railroads, storage and support structures, and a living area. This infrastructure was built to launch spacecraft using the Start-1 launch vehicle. As of 1 January 2001 three spacecraft had been launched from Svobodnyy, including the Russian communications satellite Zeya (March 1997), the American Early Bird-1 (December 1997) and the Israeli EROS-A1 (December 2000). The spaceport has five launch silos developed for RS-18 missiles, which are to be refitted for Rokot and Strela launch vehicles. Technical complexes plan to be built for preparing spacecraft.

Kapustin Yar Spaceport

Kapustin Yar is located in the Volgograd Oblast on a site whose geographic coordinates are 49° North latitude and 46° East longitude. This spaceport has a developed infrastructure, including launch and technical complexes, and telemetry stations for receiving data from launch vehicles and spacecraft at powered trajectory. At first, silo launch facilities for testing ballistic missiles were adapted for launching spacecraft. Later, the Voskhod space rocket complex, which began operation in 1973, was used.

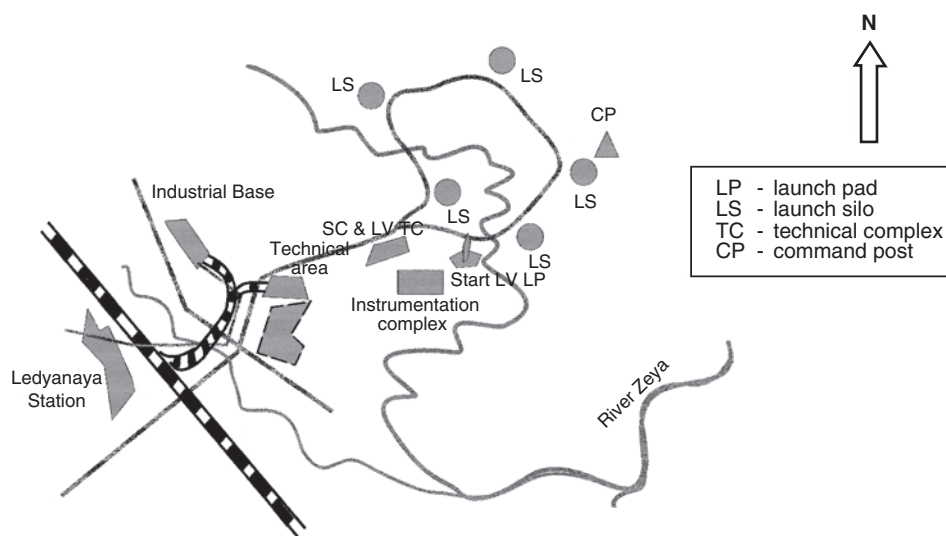


Figure 4. Schematic map of the Svobodnyy spaceport. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

This spaceport made a significant contribution to scientific research and meteorological and geophysical studies of the upper atmosphere in the framework of USSR Academy of Science programs, and also to international collaboration. In addition to facilitating development of the orbital spacecraft within the Soviet reusable space system, from 1980–1988 it was used for flight tests of the Bor-4 and Bor-5 air vehicles.

International collaboration started in 1969 to develop the technology spacecraft DS-YZ-IK, that was renamed Interkosmos-1 after injection into orbit. There were 14 launches in the framework of the Interkosmos program. In addition, spacecraft have been launched as part of joint programs with India and France. Approximately 90 spacecraft were launched from Kapustin Yar between 1962 and 2000.

Future Prospects for Development of Russian Spaceports

Future prospects for the general development of spaceports will be determined by their geographic location and their national ownership, the potential of existing launch vehicles and new improved vehicles in the future. The need to accomplish the required number of spacecraft launches, considering geopolitical and economic changes expected in the future, predetermined the use of Russian spaceports. Baikonur will be used for Russia's Federal Space Program and international and commercial space projects. Plesetsk is intended for launching spacecraft for defense, scientific, social and economic purposes, as well as for international and commercial space projects. Svobodnyy will be used for launching defense spacecraft and spacecraft for commercial space projects. Thus, Baikonur requires the implementation of measures to maintain its facilities in working order to prepare and launch spacecraft and updating of advanced launch and technical complexes taking into account future improvements of launch vehicles and spacecraft. Plesetsk requires, first of all, construction of a space rocket complex to support preparation and launch of Angara family vehicles, and improvement of the entire spaceport infrastructure. Svobodnyy needs to rebuild existing facilities and construct new ones considering the tasks that will be assigned to this spaceport.

Development of spaceports in the future will be directed at

- maintaining the high technical readiness and capacity of their infrastructure;
- raising the level of automated processes involved in preparing and launching space rockets;
- standardizing facilities for preparing launch vehicles, upper stages and spacecraft, and launching rockets;
- ensuring high reliability and safety of work performed to prepare and launch rockets;
- decreasing environmental harmful impact and operational costs.

BIBLIOGRAPHY

1. Russian Federation Law, Regarding activities in space, Moscow, 1998.
2. White paper on the Russian Federation's national space policy, Moscow, 1996.
3. Glushko, V.P., et al. *Cosmonautics Encyclopedia*. Moscow, 1985.

4. The Baikonur Spaceport. Photoalbum, Moscow, 2000.
5. Umanskiy, S.P. *Launch vehicles. Spaceports*. Moscow, 2001.
6. Stromskiy, I.V. *Spaceports of the World*. Moscow, 1996.
7. Fifty years ahead of its time. Russian Space Agency, Moscow, 1998.
8. Results of Russia's space activity in the interests of science, technology, various branches of the economy and international collaboration, Government report. Russian Space Agency, Moscow, 2001.
9. Markov, A.V. Report at the international symposium on the history of aviation and cosmonautics, Moscow, 2001.
10. Menshikov, V.A. *Baikonur. My Pain and My Love*. Moscow, 1994.
11. Kuznetsov, A.N. Potential of Russian ground-based space infrastructure and launch facilities in implementation of international space programs on the threshold of the new millennium. Report of the Deputy General Director of the Russian Aviation and Space Agency, Moscow, 2001.
12. Russian cosmonautics on the boundary between centuries, Proc. Moscow space club, number 6, Moscow, 2000.
13. Major Trends in Russia's Activities in Space, Russian Cosmonautics. Moscow, 1997.
14. Russia's rocket and space industry, Catalogue of business, organizations, and institutions. Russian Aviation and Space Agency, 1999–2000.
15. Kiselev, A.I., et al. Cosmonautics on the boundary between millennia. *Conclusions and prospects*. Mashinostroyeniye, Moscow, 2001.
16. News of Cosmonautics: Nos. 21/22 (1998), No. 3 (1999), No. 6 (1999), No. 9/99 (1999); Nos. 3, 4, 10 (2000); Nos. 1, 2, 5, 6 (2001).

ALEXANDER N. KUZNETSOV
ALEXANDER F. DEDUS
Russian Aviation and Space Agency
Russia

RUSSIA'S LAUNCH VEHICLES

The history of cosmonautics and rocket building in Russia may be considered to have begun at the start of the twentieth century. The founder of these disciplines was the great Russian scientist K.E. Tsiolkovsky; major theoretical contributions were made by F.A. Tsander and Yu. V. Kondratyuk. Even in the prewar years, experimental work was being conducted to develop rocket technology for a number of purposes. In 1921, the Gas Dynamics Laboratory (GDL) was founded in Moscow, and in the early 1930s, the Group for the Study of Reaction Propulsion (GIRD) was formed. In 1933, the Reaction Scientific Research Institute (RNII) was created from the GDL, and Moscow GIRD, and in the late 1930s, Design Bureau (KB-7) was founded. All of this resulted in the development of the world's best volley fire (Katyusha) and aircraft missile systems. At the same time, experimental, jet aircraft and liquid-fueled guided missiles were being designed and developed. The production facilities and supporting fundamental and applied science for these areas developed rapidly, serving as the basis for the start of the powerful space and rocket industry in the 1950s.

S.P. Korolev, the director of Special Design Bureau No. 1 (OKB-1) is considered to be the founder of practical cosmonautics and rocket building. He was the Chief Designer of the first Russian rockets, including the first intercontinental ballistic missile, R-7, which was the base rocket for the development of the earliest space launch vehicles (LV). The country's success in developing rocket technology, in discoveries of the space era and the conquest of space are also associated with the names of the following outstanding designers: V.P. Glushko, N.A. Pilyugin, M.S. Ryazanskiy, V.N. Chelomey, V.I. Kuznetsov, V.P. Barmin, M.K. Yangel, V.F. Utkin, G.N. Babakin, V.S. Budnik, A.M. Isayev, S.A. Kosberg, V.P. Makeyev, M.F. Reshetnev, V.P. Mishin, V.N. Solovyev, A.D. Nadiradze, and V.M. Kovtunenkov; the scientists: M.V. Keldysh and B.N. Petrov; the military specialists: V.I. Voznyuk, A.G. Karas, A.I. Sokolov, and A.G. Mrykin; directors and managers: K.N. Rudnev, D.F. Ustinov, S.A. Afanasyev, L.V. Smirnov, G.A. Tyulina, Yu.A. Mozzhorin and many others.

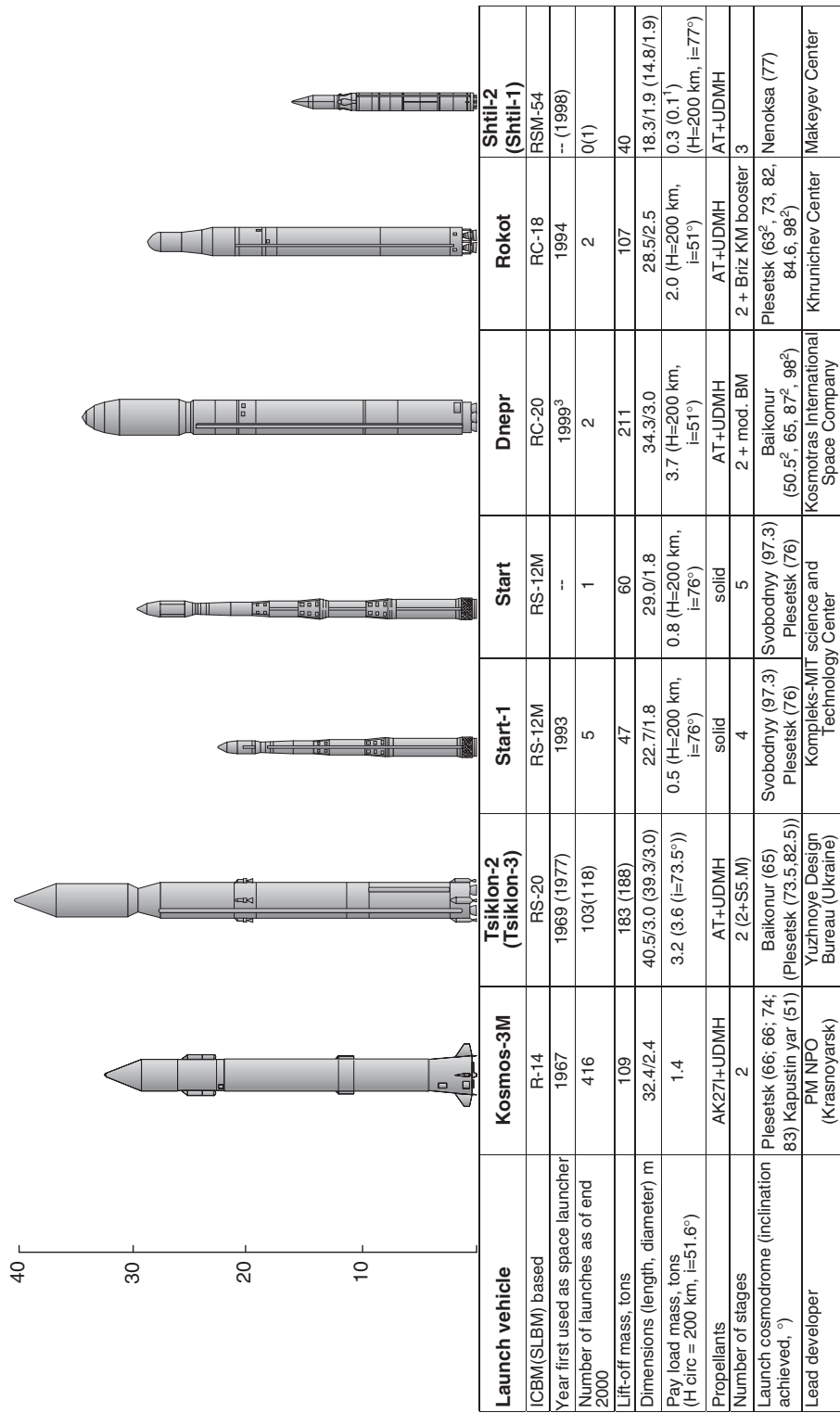
In addition to OKB-1 (which is currently Korolev Rocket and Space Corporation Energia (Korolev, Moscow region) such well-known enterprises as Khrunichev State Research and Production Space Center (Moscow), Progress State Research and Production Rocket Space Center (Samara), Makeyev Design Bureau (State Rocket Center) (Miass), Polyot Production Association (Omsk), Research and Production Association for Machine Building (Reutov), Lavochkin Research and Production Association, Yangel State Design Bureau Yuzhnoe (Dnepropetrovsk, Ukraine) and many others were engaged in activity related to rocket and space technology.

At present time, the development of space rocket building in Russia is associated with directors and chief designers such as Yu.N. Koptev, Yu.P. Semenov, D.I. Kozlov, A.A. Medvedev, S.D. Kulikov, G.A. Yefremov, Yu.S. Solomonov, V.G. Degtyar and others.

The current Russian system of launch vehicles evolved during the process of formulating and addressing a whole set of tasks in the interests of various users, employing the experience accrued through design and use of previously developed military rockets and space complexes. The available capacities of existing launch vehicles, in general, satisfies the requirements of the spacecraft that are in current use and planned for the near future for injection into circular and elliptical orbits of various altitudes and inclinations and interplanetary flight trajectories. Spacecraft are generally launched from the Baikonur and Plesetsk cosmodromes, whose infrastructures were developed based on progressive evolution of space activity. A few light launch vehicles are launched from the Svobodnyy cosmodrome and Kapustin Yar range. Current space rocket systems have a high level of reliability and cost efficiency.

The current launch vehicles includes the expendable launch vehicles Kosmos-3M, Tsiklon-2,3, Molniya-M, Soyuz-U, Zenit-2, Proton-K and a number of launch vehicles developed as part of the military rocket conversion program. The launch vehicles used can be divided into light, middle, and heavy classes depending on the payloads launched.

The light class includes the Kosmos-3M, Tsiklon-2, 3 and also the converted Dnepr, Rokot, Strela, Start, Start-1, and Shtil launch vehicles (Fig. 1); the mass of payloads inserted by these rockets into low orbits is 0.3–3.7 tons. All launch vehicles, aside from the solid-fuel rockets of the Start type, use AT (nitrogen



1) spacecraft mass limited by small size of payload module; 2) opening of routes being investigated 3) launcher prototype

Figure 1. Light-class launch vehicles. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

tetroxide), AK-27I (nitric acid + 27% N_2O_4 + iodine) + UDMH (unsymmetrical dimethyl hydrazine) propellants.

The prototype launch vehicle Kosmos-3M (RN 65S3) was developed by the Yuzhnoye Design Bureau (Dnepropetrovsk, Ukraine) on the basis of the intermediate ballistic missile R-14 (SS-5). Kosmos-3M is the result of a number of updates of the prototype vehicle and has been used since 1967. Since 1970, it has been serially produced by the Polyot Production Association (Omsk).

The two-stage Kosmos-3M launch vehicle can inject spacecraft into orbits up to 1700 km high by twice firing the second-stage liquid fuel sustainer engine and using a low thrust system. This launch vehicle was used to launch light spacecraft as part of Federal programs, and starting in the late 1990s, was also used to launch commercial spacecraft. At present, Kosmos-3M is no longer produced.

The two- and three-stage Tsiklon-2 and Tsiklon-3 launch vehicles were developed by the Yuzhnoye Design Bureau in 1969 and 1977, respectively. Both modifications were based on the R-36 (SS-9) military missile. The first two stages of these vehicles are virtually standardized. As its third stage, Tsiklon-3 uses the S5.M unit and repeatedly fires the sustainer engine, making it possible to expand the range of injection orbits. Production of the main components of the Tsiklon was handled by the Yuzhmashzavod Production Association (Dnepropetrovsk, Ukraine). At the present time the production of these rockets has been discontinued.

The Tsiklon System is highly automated and does not require personnel at the launch facility between the time the rocket is delivered to it and the moment of launch.

In the early 1990s, development work started on launch vehicles as part of the program to convert intercontinental ballistic missiles and submarine-launched ballistic missiles that were to be eliminated in accordance with the arms reduction treaty. Solid-fuel launch vehicles of the Start class were developed strategically by the Kompleks-MIT Science and Technology Center based on the RS-12 M (SS-25) intercontinental ballistic missile. These vehicles are intended for launching small spacecraft with various uses into low near-Earth orbit. The first launch of the four-stage Start-1 vehicle occurred at the Plesetsk Cosmodrome in March 1992. Further launches of this vehicle as part of Federal and commercial programs took place at the Svobodnyy Cosmodrome. Development of a five-stage version of Start is currently underway.

The Dnepr launch vehicle was developed by the Kosmostras International Space Company and was based on the most powerful intercontinental ballistic missile in the world, RS-20 (SS-18). This liquid-propellant, two-stage ICBM was developed in 1973 by the Yuzhnoye Design Bureau and produced serially at the Yuzhmashzavod Production Association. There were two successful launches of the Dnepr prototype carrying foreign commercial spacecraft in April 1999 and September 2000. These launches occurred at the Baikonur cosmodrome from the silo launch facility. All of the main components of this launch vehicle are standard and available without modification. There are proposals to equip the launch vehicle with a more powerful third stage (Dnepr-M).

The Rokot launch vehicle, which is distinguished by its third stage, was based on the RS-18 (SS-19) ICBM. For its third stage, Rokot (chief developer Khrunichev Center) uses the Briz-KM upper stage and multiple firings of the

sustainer engine, making possible various injection patterns. The first launch of Rokot took place in December 1994 from the ICBM RS-18 silo at Baikonur and injected a spacecraft into orbit. In May 2000, Rokot was launched from Plesetsk carrying two mock-ups of the Iridium spacecraft. The launch complex for Rokot launches was built at this cosmodrome by rebuilding the existing facility for launching Kosmos-3M and the technical complex for preparing Rokot and its spacecraft Tsiklon-3 technical facility.

As part of the program for converting submarine-launched missiles, the Makeyev Design Bureau Center undertook to adapt certain rockets for use as launch vehicles. In July 1998, two Tubsat spacecraft were launched into low near-Earth orbit using the Shtil-1 launch vehicle (a conversion of the RSM-54 submarine-launched missile) from an underwater position in the region of the Barents Sea. The standard three-stage, submarine-launched ballistic missile was adapted for spacecraft launch: a special frame to hold the spacecraft was installed and the flight program was altered. In addition, a special container with telemetry instrumentation that allowed ground control to monitor the injection was mounted on the third stage. The State Rocket Center has made a proposal to update Shtil. The more powerful Shtil-2 and Shtil-3 are proposed for launching small low-orbit spacecraft.

The intermediate class includes launch vehicles using an oxygen-kerosene propellant: Molniya-M, Soyuz-U, and Zenit-2 (Fig. 2). The range of payloads they can inject into low orbits is 6.8–13.7 tons. Molniya and Soyuz launch vehicles were developed by the team at OKB-1 under the direction of S.P. Korolev based on R-7 ICBMs. Note that R-7 ICBMs (R-7A, S-6) are the base rocket for development of a whole series of modified space launch vehicles, in particular,

- the Sputnik launch vehicle, which, in October 1957 launched the first artificial Earth satellite into Earth orbit. The payload capacity of this two-stage launch vehicle in low orbit was approximately 2.0 tons.
- the Vostok launch vehicle, which was the Sputnik vehicle equipped with a third stage (block “E”) that enabled launches of flights to the Moon by the unmanned Luna-1, -2 and 3 in September–October 1959 and the April 1961 launch of the manned spacecraft Vostok with the Earth’s first cosmonaut, Yu.A. Gagarin. The payload capacity of this three-stage launch vehicle in low orbit was approximately 4.8 tons.

Starting in 1961, all of the work to develop, improve, flight-test, and operate R-7 rockets was assigned to the Progress State Research and Production Rocket Space Center (Samara). The first launch of the four-stage Molniya vehicle occurred in 1960. In 1964 and 1985, this launch vehicle underwent substantial updating to expand its capacities and maintenance safety. At present, the Molniya-M is being used. It has an “L” upper stage with a liquid propellant engine fired in weightlessness. This module contains the control system, which controls the flight of both modules L and I. The Molniya-M is designed to launch spacecraft for interplanetary flights and into highly elliptical orbits. It has been used for launches of spacecraft of the Luna, Venera, Mars, Molniya, and Prognoz type spacecraft.

The three-stage Soyuz-U is an updated version of the Soyuz launch vehicle, used from 1966–1973. It can launch various kinds of spacecraft, including

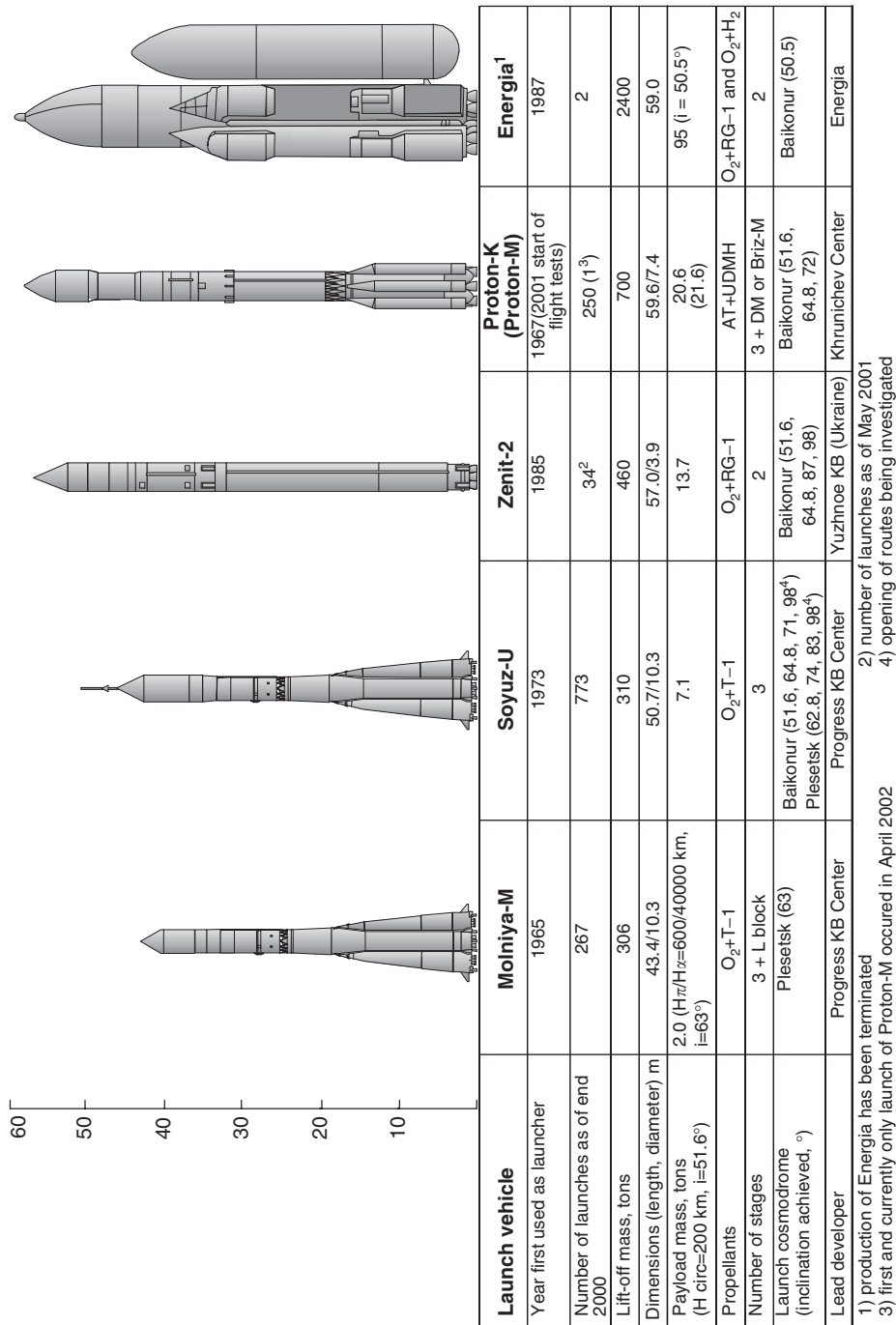


Figure 2. Intermediate- and heavy-class launch vehicles. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

manned spacecraft from the Baikonur Cosmodrome and unmanned craft from Plesetsk. Between 1982 and 1995, a modification of Soyuz-U, called Soyuz-U2, was used for the manned space programs. Unlike Soyuz-U, Soyuz-U2 used the "sintin" propellant instead of kerosene T-1 in its core module. At present, this launch vehicle is no longer manufactured.

At present, the Progress State Research and Production Rocket Space Center (Samara), is conducting work on a phased modernization of Soyuz launch vehicles. The tasks of modernizing Soyuz involve the phased development of a standardized three-stage launch vehicle, Soyuz-2 (Fig. 3).

During the initial phase, to allow launch of Soyuz-TMA spacecraft and the cargo spacecraft Progress-MA, work is being conducted to provide a minor update of the Soyuz launch vehicle. The essence of this update is the use of updated stage I and II engines on the basic Soyuz vehicle, and minimal changes of the existing control system, the tank drainage and synchronization, and indicating speed regulation systems. This launch vehicle, designated Soyuz-FG, using a nose cone 3.0 m in diameter, can inject a payload weighing up to 7.4 tons into low Earth orbit. Flight tests of Soyuz-FG began in 2001.

The next phase, along with the use of launch vehicle engines on stages I and II, will involve

- use of a control system based on a highly efficient digital computer that has modern components and advanced software;
- use of a new digital radiotelemetry system.

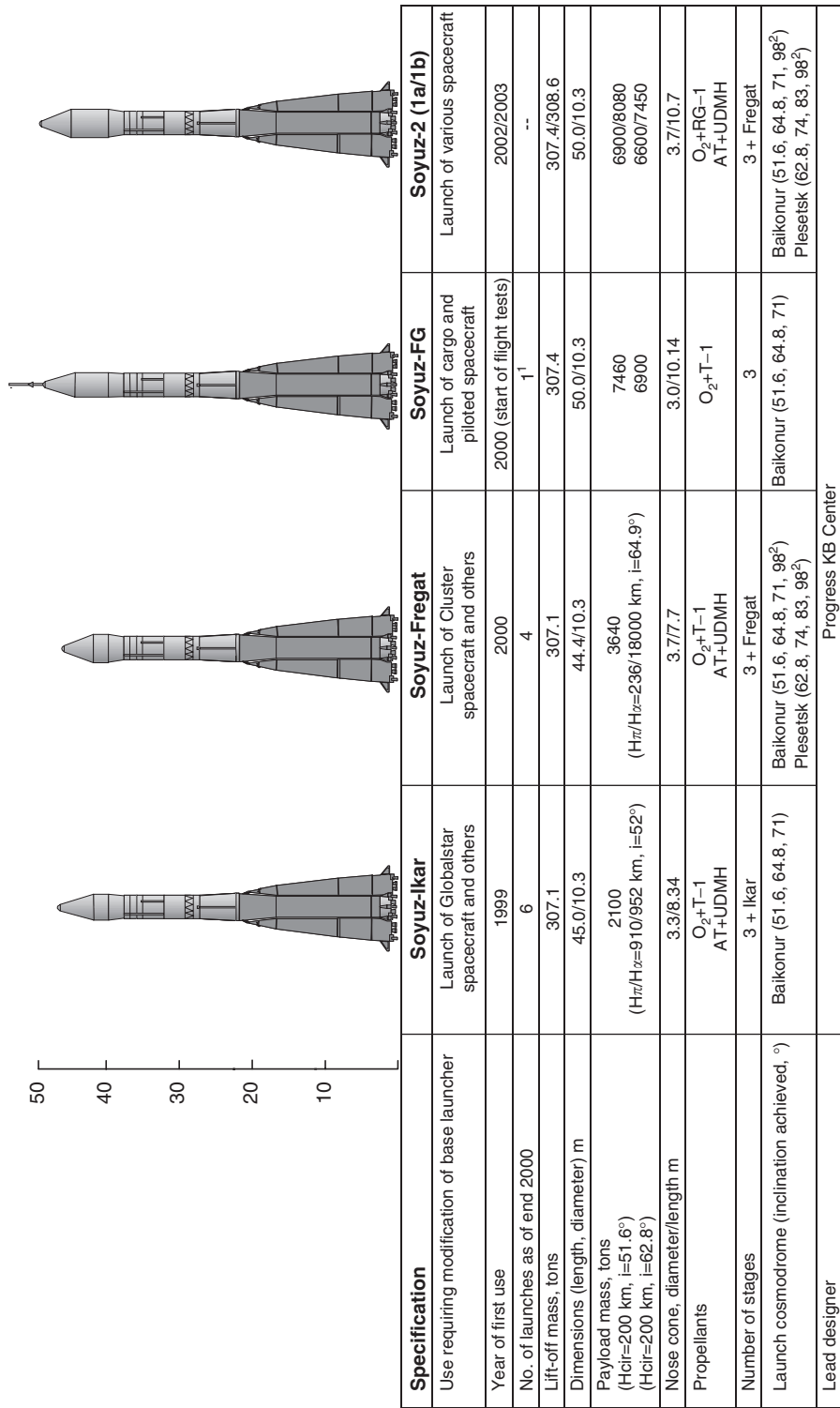
The new control system will provide very precise injection into orbit and also will increase the diameter the nose fairing from 3.3 to 4.1 m. The Soyuz-2 launch vehicle will have a payload capacity for injection into low Earth orbits of up to 6.9 tons.

During the last phase, the Soyuz-2 launch vehicle will be equipped with a more powerful stage III and a new liquid propellant engine developed by the Design Bureau for Chemical Automation, which will increase the payload mass for low near-Earth orbit to 8.1 tons.

To deploy Globalstar telecommunications satellites, the Starsem Company (a Russian-French joint venture), used the Soyuz-Ikar system. The Ikar upper stage was developed at the Progress Center using as a base the equipment module of the Kometa spacecraft. Since 1999, there have been six successful launches of Soyuz-Ikar.

To expand the altitude range for spacecraft orbital injection (right up to geostationary orbit), the Lavochkin Research and Production Association developed the upper stage Fregat, based on the propulsion unit of the Fobos spacecraft, for use with Soyuz launch vehicles (Soyuz-Fregat launch vehicle). At present, the certification of the system has been completed, so that it is ready for use. In 2000–2001, four successful launches of this vehicle were completed as part of the Cluster program. The Soyuz-Fregat can insert a payload with a mass of 3640 kg into highly elliptical orbits (for $H_{\pi}/H_{\alpha} = 236/18,000$, $i = 64.9^{\circ}$).

The Starsem Company has commissioned development of a commercial variant of this launch vehicle, under the name Soyuz-ST. The major difference between this vehicle and Soyuz-2 is a larger nose cone based on the nose fairing



1) number of launches as of July 2001

2) opening of routes being investigated

Figure 3. Specification of Soyuz type launch vehicle modifications. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

from the Ariane-4 launch vehicle (diameter 4.1-m) to increase the space available for the payload. This launch vehicle can be launched from the Baikonur and Kourou cosmodromes.

The two-stage launch vehicle Zenit-2, developed by the Yuzhnoye Design Bureau and manufactured in Ukraine, is intended for injecting midclass unpiloted spacecraft into low and intermediate altitude circular orbits (including sun-synchronous orbits). This system requires almost no manual labor for service, when it is being prepared for launch and when it is removed from the launch facility if the launch has been aborted. When launched from Baikonur, Zenit-2 injects payloads of up to 13.7 tons into low orbits.

The Zenit and the DM upper stage (developed by RSC Energia), based on advanced innovative technology, formed the basis for the international commercial Sea Launch project. When launched from the equatorial zone, Zenit-3SL with the DM-SL upper stage can inject a payload weighing 2.9 tons into geostationary orbit. In 1999, the first demonstration launch of this rocket took place from the Odyssey floating platform in the vicinity of Christmas Island. After that, Sea Launch began commercial operations. In mid-2001, there were seven launches of Zenit-3 SL; one of which was unsuccessful.

Work has started to develop the next generation of intermediate-class launch vehicles. RSC Energia, with foreign investment, is developing the Aurora midclass launch vehicle. The Aurora is planned for launches of spacecraft for Federal programs and commercial customers. For commercial launches, the Asia Pacific Space Center (APSC) has commissioned a cosmodrome on Christmas Island (Australia). According to the draft project published by RSC Energia in 2001, the launcher will deliver payloads with a mass of 12.0 tons to low circular orbit and 2.1 tons to geostationary orbit, when the launch occurs from the cosmodrome on Christmas Island, and payloads of 11.0 and 0.9 tons, respectively, for launches from Baikonur.

The heavy class of launch vehicles includes Proton-K, developed by the Khrunichev Center using AT + UDMH propellants (Fig. 2). Two versions of this launcher are used: a three-stage for delivering spacecraft into low orbits (Mir modules and International Space Station modules, and heavy spacecraft) and the four-stage version with the DM transfer stage to inject spacecraft into high-energy orbits (including geostationary transfer orbits, geostationary orbits, and interplanetary trajectories).

The Proton-K has earned a reputation as the most reliable (flight reliability = 0.97), well-tested (230 launches have been performed), and cost efficient launch vehicle of the heavy class. Progressive design changes incorporated as it was built have allowed it to keep pace with growing demands during almost three decades. At present, its payload capacity for orbits 200 km high and inclination of 51.6° is 20.6 tons, and its payload capacity in geostationary orbit has been increased to 2.6 tons.

The Khrunichev Center is modernizing the Proton-K. The development of the Proton-M launcher is being conducted to replace the outmoded analog control system by a modern system with an onboard digital computer and a flexible injection program; to reduce environmental pollution; to provide a smaller impact zone for the separated elements; and also to improve energy performance (in low orbit, up to 21.6 tons and in geostationary orbit using a DM or Briz-M upper stage up to 3.0 tons). In April 2001, the first launch of Proton-M took place,

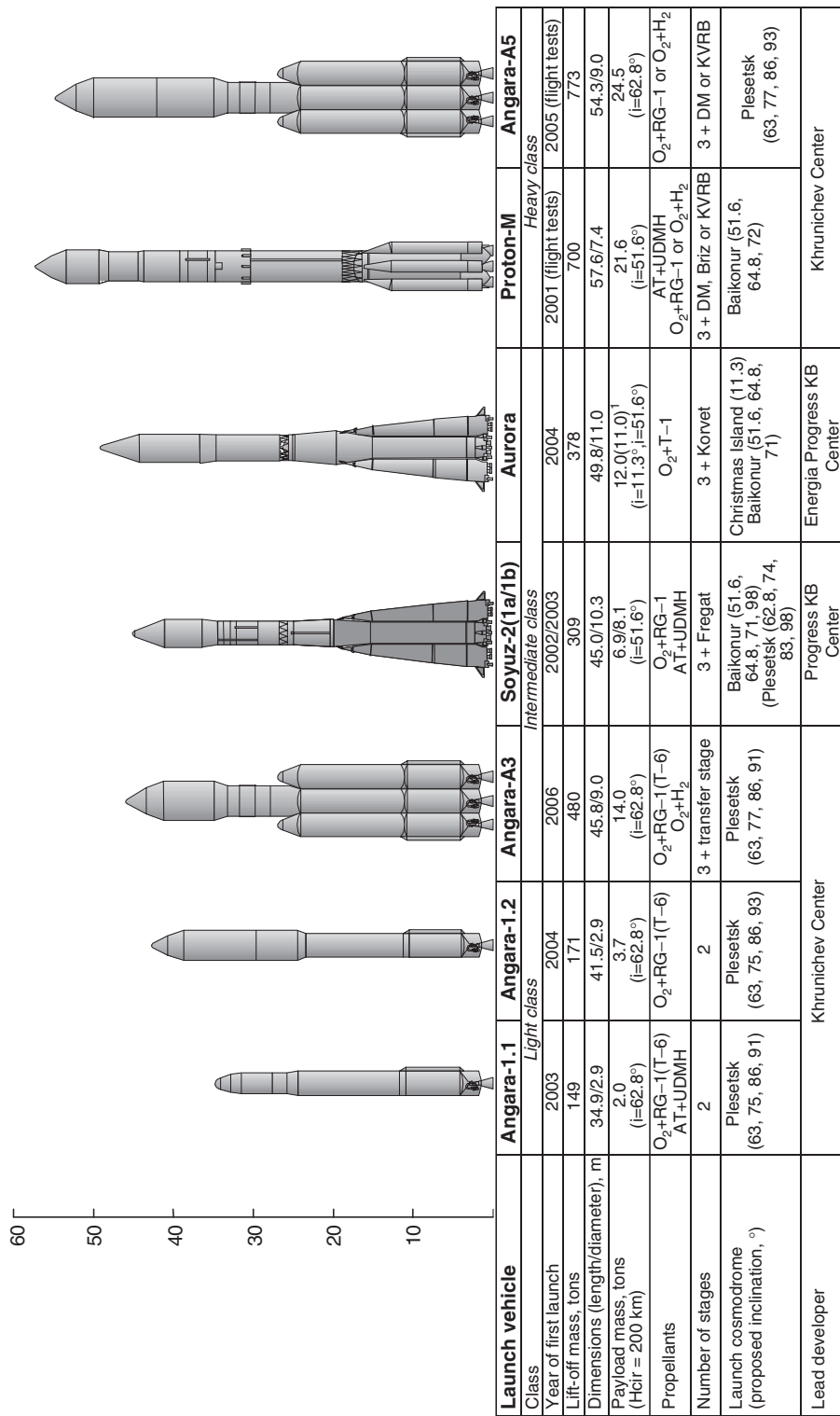
starting its flight-test phase. In the future, plans call for using an oxygen–hydrogen upper stage, which will enable an increase in payload mass to geostationary orbit to 4.0 tons.

In the 1970s, Special Design Bureau-1 (in 1966, renamed the Central Design Bureau for Experimental Machine Building) developed the superheavy N-1 launch vehicle for the manned Lunar program. The launch mass of the rocket was approximately 2800 tons, and the length of the rocket with payload was approximately 100 m. Propellants were supercooled oxygen and kerosene. The N-1 launcher was supposed to inject a payload of 95 tons into a parking orbit of 220 km. The first test launch of the rocket occurred in 1969. There was a total of four launches. Unfortunately, all launches failed and in 1974, it was decided to discontinue work. Note that these N-1 developments were subsequently used to develop the Energia launcher. The Energia, which was flight-tested in the late 1980s, used oxygen and kerosene and oxygen and hydrogen as propellants and had a payload capacity for low orbit of 95.0 tons, belongs to the superheavy class (Fig. 2). After two launches (both were successful) in May 1987 and October 1988 as part of the program to create the reusable Energia-Buran space shuttle system, the production of Energia was terminated.

In the near future, the development of the Russian system of launch vehicles (Fig. 4) stipulates the use of Soyuz-2 and Proton-M launchers to implement Federal and commercial programs. The Zenit launcher may also be used to perform particular launches in the Federal program.

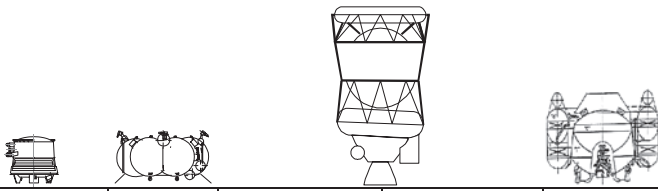
Creation of the new generation of launch vehicles is linked, first and foremost, with development of the Angara launch vehicle family (lead development at the Khrunichev Center). The main goal of the project is to develop the Angara-A5 heavy-class launcher. This launcher is being developed to provide Russia with guaranteed access to space from the Plesetsk cosmodrome (including launches into geostationary orbit) for implementing the Russian National Space Policy and future advancement of Russian launch vehicles in the global market. A modular technique for building this family of launchers of different classes has been proposed (Fig. 4). As the core module, designers are considering the new universal rocket module (URM-1), a development with an oxygen–kerosene engine, developed by NPO Energomash. This module will serve as the basis for constructing launchers of various classes, starting with light-class launchers Angara-1.1 and Angara 1.2, which will use one URM as the first stage and different second stages. Successful development of URM as a component of relatively inexpensive light-class launchers, and also the possibility of obtaining additional funds from potential commercial clients for their launches, shall become the basis for developing more powerful launchers in the Angara family. The heavy class Angara-A5 will have a cluster of five URMs and the midclass Angara A3 will use a cluster of three URMs. Launches of Angara family launchers are planned for the Plesetsk cosmodrome with maximal use of the existing launch and technical facilities at that site.

To expand the spectrum of orbits with regard to altitude and inclination, upper stage modules are used for Russian intermediate- and heavy-class launch vehicles. Figs. 5 and 6 show the designations and characteristics of current and modernized (future), upper stage modules for intermediate- and heavy-class



1) for launch from Australia's Christmas Island, in parentheses Baikonur launch

Figure 4. Future launch vehicles. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.



Upper stage	Ikar	Fregat	DM-SL	DM (11C86-01)	Briz-M
Launch vehicle	Soyuz	Soyuz	Zenit-3SL	Proton-K	
Year first used	1999	2000 (fl test)	1999	1967	1999 (fl.test)
No of launches, as of end 2000	6	4	7 ¹	232	2 ²
Propellants	AT+UDMH	AT+UDMH	O ₂ +RG-1	O ₂ +RG-1/"sintin"	AT+UDMH
Payload mass in orbit, tons	2.2 ³	0.3	2.9 (sea launch)	2.4/2.6	2.1/2.3 ⁴
Geost. orbit from Baikonur	--				
Mass of fueled stage, tons	3.3	6.5	18.5	18.2	22.4
Stage dimensions (DxL)m	2.7 × 2.6	3.3 × 2.6	3.7 × 6.3	3.7 × 6.3	4.1 × 3.1
Mass of expendables, tons	0.9	5.4	15.0	15.0	5.2/14.7 ⁵
Total trust of cruise engine, tons sec	0.3	2.0	8.5	8.5	2.0
Specific impulse, kg x s/kg	324	327	353	352/361	325.5
Developer	Progress KB Center	Lavochkin NPO	Energia	Energia	Khrunichev Center

1) number of launches as of May 2001

2) number of launches counting the first launch of Proton-M in April 2001

3) on launch into mid-near Earth orbit ($H_{\pi}/H_{\alpha}=906/948\text{km}$, $i=52^{\circ}$)

4) numerator-direct injection into geostationary orbit, denominator- 10 hour mode (with intermediate orbital phasing)

5) numerator-central module, denominator-auxiliary propellant tank ejected in flight

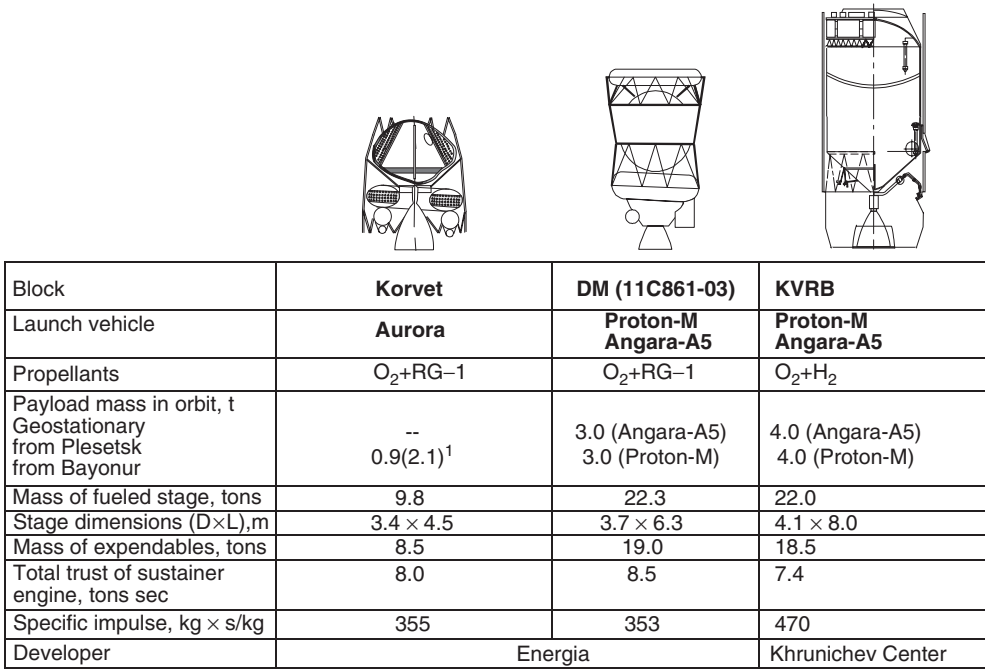
Figure 5. Upper stages used for intermediate- and heavy-class launch vehicles. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

launchers. Depending on the propellant used, these, modules can be divided into three groups:

- upper stage modules using storable toxic propellants (Ikar, Fregat, Briz-M);
- upper stage modules using oxygen-kerosene propellants (block DM type);
- upper stage modules using oxygen and hydrogen propellants.

Further improvement of the lift capacities of heavy- and intermediate-class launchers is associated with the development and use of transport modules with low thrust engines, particularly, electric rocket engines (ERE). The current technical level of development and cumulative flight experience in the application of ERE and solar cells has created the preconditions for developing sustainer engine systems for energy-intensive transport operations in near-Earth space. The use of low thrust engines with substantially better efficiency than traditional liquid propellant engines will require implementing new patterns of injection into high-energy orbits, including geostationary orbits. The time required for orbital injection would be increased from several hours to several months, which demands the development of a spacecraft adapted to such prolonged injection.

The main line of development of space launch systems in the twenty-first century, first and foremost, is based on the gradual shift to reusable space transportation systems, including space-rocket and air-space systems. As



1) in parentheses are data for launch from Christmas Island cosmodrome

Figure 6. Modernized and future upper stages for intermediate- and heavy-class launch vehicles.

research has shown, the shift to reusable launch vehicles of a new generation can:

1. decrease the cost of launching a unit of payload mass compared with the traditional expendable launch vehicles:
 - by a factor of 1.5–2 if reusable space transport systems are built with the optimum percentage of reusable components using a level of technology that will be available in the near future;
 - by a factor of 5 to 7 if a completely reusable vehicle were built using high level advanced technology.
2. increase the mission performance probability and crew safety by a factor of at least 5 compared to the level associated with current launchers.
3. virtual elimination of environmental pollution by maximum decrease, up to complete elimination, of the hazard zones along the launch trajectories, as a result of use of nontoxic propellants, and also injection of spacecraft into orbit without accompanying debris of spent launcher parts.

At present, more than 15 Russian scientific, research, and design organizations are working on a program to develop the scientific and technical requirements for a reusable transport system. Results of a comparative analysis of various vertical and horizontal launch reusable transport vehicle designs allowed to

select four reusable space transportation system versions. Two of them are based on technology available in the near future:

- A system based on a two-stage reusable all-azimuth rocket with a vertical launch and multiple return (to the launch site) rocket booster in the first stage and an expendable injection module as the second stage. The payload in this design could be a spacecraft under the nose fairing, if injection is to take place according to the traditional pattern, or an orbital transport spacecraft, consisting of a reusable orbital spacecraft and an expendable cargo module, for performing transport and technical maintenance operations. At present, the Khrunichev Center in cooperation with the NPO Molniya, is developing a reusable module of the first stage Baykal, which could be used in an all-azimuth rocket launch vehicle in the near future (Fig. 7). The reusable Baykal module has a landing mass of about 18.5 tons, a length of the order of 29 m, and a diameter of 2.9 m. It has the configuration of a high-wing monoplane with a rotating wing attachment above the body of the fuselage. An RD-191 M liquid propellant engine is installed in the Baykal tail section. In the nose, there is an RD-33 jet engine unit for cruise flight on return to the launch area and landing at an airport. A full size technological mock up of the Baykal was displayed in June 2001 at an aerospace exhibit at Le Bourget and in August 2001 at the MAKS-2001 airshow.
- A two-stage reusable aerospace vehicle system with horizontal launch, consisting of an An-225 subsonic carrier aircraft and a reusable orbital aircraft with an expendable external propellant tank.

Two versions were selected based on more advanced future technology:

- a single stage fully reusable space rocket plane with vertical launch and horizontal landing;
- a two-stage reusable aerospace system with horizontal launch, consisting of a hypersonic aircraft booster with a combined propulsion system and a rocket orbital boost stage.

The development of a reusable space transportation system is linked to the solution of a whole series of technical problems; the key ones are

- development of reusable sustainer liquid propellant engines;
- development of reusable returnable booster rockets for the first stage;
- development of materials, designs, technologies, and components for decreasing the mass of the assemblies and systems by 30–60% and more, compared to the current level;
- development of methods for minimizing work entailed in turnaround servicing, including methods of monitoring and diagnosing the postflight status of equipment, including the large cryogenic containers;
- development of technical principles and methods for providing a qualitatively new level of safety and preservation of equipment in emergency situations.

Thus, proposals call for creating a new generation of space transportation systems by 2010–2015, which, because of the use of advanced engineering

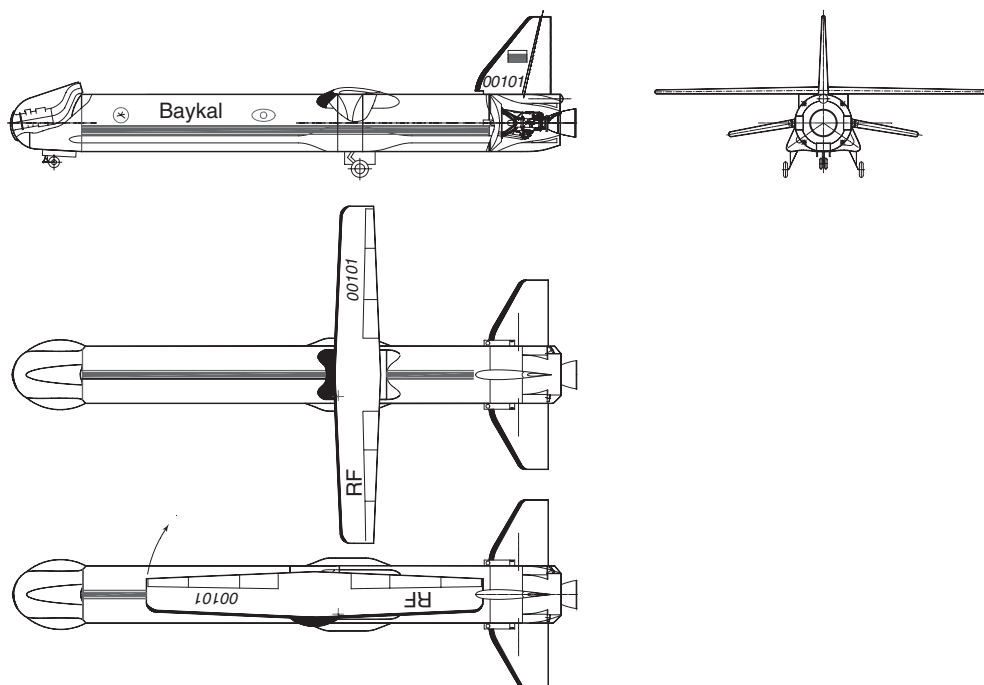


Figure 7. First stage of the reusable Baykal module. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

designs and technologies (reusability, advanced materials, designs, components propulsion units, control systems, ground servicing technologies, etc.) will make it possible (by a factor of 2–4) to decrease annual expenditures significantly on transportation support, expand the consumer potential of launch vehicles, improve their performance characteristics (reliability, safety, readiness for launch, and others) and make Russian transportation systems competitive on the launch vehicle world market.

BIBLIOGRAPHY

1. Umanskiy, S.P. *Launch Vehicles and Cosmodromes*, edited by Yu.N. Koptev Director of the Russian Aerospace Agency. Redstart +, Moscow, 2001.
2. Kiselev, A.I., A.A. Medvedev, and V.A. Menshikov. *Cosmonautics on the Threshold of the Millennium. Conclusions and Prospects*. Mashinostroyeniye-Polet, Moscow, 2001.
3. Karpenko, A.V., A.F. Utkin, and A.D. Popov. *Russian Strategic Rocket Systems*. Nevskiy Bastion, St. Petersburg, 1999.
4. *Novosti kosmonautiki* (News of Cosmonautics), No. 1–12, 2000, No. 1–7 (2001).

ALEXANDER N. KUZNETSOV
Russian Aviation and Space Agency
Russia

S

SATURN SYSTEM

Introduction

Saturn has intrigued human beings ever since they noticed that it belongs to the group of objects called *planets*, which move relative to the “fixed stars” in the sky. Saturn is the slowest moving of the ancient planets. It takes fully 29.46 years to return to the same position in the sky. Two orbits of Saturn are thus very close to an integral number of 59 Earth years, so that the phenomena of Saturn repeat on roughly the same yearly date after an interval of 59 years. This was recognized by the ancient Babylonians and Greeks. Thus, Saturn was associated with the measurement of time being named *Cronos* (time) in Greek legend and serving as the father of the lord of planets Jupiter (Zeus). The Romans celebrated the winter feast *Saturnalia* in his honor, and the name Saturn today is fixed in our language in the word *Saturday*.

While all of the major planets have faint ring systems, the rings of Saturn are so bright and extensive that Saturn is undoubtedly the most spectacular telescopic object in the sky. When viewed with a good telescope, Saturn seems to float mysteriously and serenely in the vast reaches of space. When the rings of Saturn are opened to their maximum, they provide 1.3 times as much light as the ball of the planet. During ring plane crossings when the rings are edge-on to an observer on Earth and thus become invisible, the visual brightness of the Saturn system diminishes by 0.92 magnitudes. During opposition, when the Sun, Earth, and Saturn are all lined up with Saturn being on the nightside of the Earth, the ball of the planet has a diameter of 19.4 arcsec, but the major visible rings (A, B, and C) add another 24.4 arcsec to its extent, a total of 43.8 arcsec. This is about the size of Jupiter, even though that planet is bigger and much closer to us.

Viewing Saturn with his primitive telescope, Galileo, in 1610, was the first to notice that Saturn was odd. He thought that Saturn was composed of three, “I have observed the most distant planet to be a triple one”, he wrote. Other

observers thought Saturn had “handles” or ansae; a name that is still applied today to the ring system of Saturn (east or west ansae). It was not until 45 years later in 1655 that the versatile Dutch scientist Christiaan Huygens solved the problem of Saturn’s rings. Huygens also discovered Saturn’s largest and brightest satellite, Titan, a satellite so intriguing that we devote a whole section to it. Incidentally, Huygens felt that the solar system was now complete and no more objects could be discovered; because with six planets and six satellites known, a perfect number of 12 had been attained. These metaphysical arguments were soon dispelled by the discovery of four more satellites of Saturn by the Italian/French astronomer, Jean Dominique Cassini: Iapetus (in 1671), Rhea (in 1672), and Dione and Tethys (both in 1684). In 1675, Cassini also discovered the prominent division between the A and B rings, called the Cassini division.

For the next 250 years or so, until the end of the nineteenth century, visual telescopic observation was the tool of the astronomer. Progress was slow but steady, and work was conducted on such subjects as the dynamics of the solar system and the best parameters for Saturn’s orbit, the orbits of its satellites, the dimensions and changing aspect of its rings, the mass and size of Saturn, its shape or its oblateness, its density and the search for features such as belts or zones in Saturn’s atmosphere or transient spots which allowed determin of Saturn’s rotation rate. Most of the well-known astronomers of that period contributed to our understanding of Saturn: William Herschel, Friedrich Bessel, William Bond, Johann Encke, Wilhelm Struve, William Lassell, Edward Barnard etc.

Our understanding of the Saturn system accelerated considerably at the beginning of the twentieth century as the new era of science and technology developed. Some of the technological advances that made this possible were the invention of the photographic process, the development of spectroscopy, electronic photometry, high-speed computers, and modern electronics that use large-area, high quantum efficiency detectors such as CCDs (charge-coupled devices). Advances in technology culminated in humanity’s age old dream of actually visiting the planets using spacecraft. A summary of Saturn spacecraft exploration is given in Table 1. The Pioneer spacecraft, as its name implies, was used as a

Table 1. **Spacecraft Missions to Saturn**

Name	Mass	Launch	Saturn encounter	Comments
Pioneer 11	258 kg	05 April 1973	01 Sept. 1979	Also Pioneer Saturn; first spacecraft to reach Saturn
Voyager 1	815 kg	01 Sept. 1977	13 Nov. 1980	
Voyager 2	815 kg	20 Aug. 1977	27 Aug. 1981	
Cassini/Huygens	5650 kg	06 Oct. 1997	July 2004	ESA/NASA mission: Saturn orbiter and Titan probe
Huygens probe	318 kg		Nov. 2004	Descent probe into Titan’s atmosphere

pathfinder to test the space between Earth and the major planets and to investigate the environment around Saturn. This was followed by the much more complex and sophisticated spacecraft, Voyager 1 and 2, which visited both Jupiter and Saturn and went on to explore the Uranus and Neptune systems. The Voyager spacecraft helped tremendously in our understanding of Saturn and sent back hundreds of close-up pictures that were essentially unimaginable to scientists and philosophers, even a mere hundred years ago. To put this achievement into proper perspective, one must imagine what Galileo or Huygens would have exclaimed, had they had been presented with the beautiful close-up spacecraft images of Saturn's globe, its rings, and satellites. A newer and even more sophisticated spacecraft called *Cassini/Huygens* is now on its way to provide us with even more information on Saturn. This is a joint venture between the U.S. National Aeronautics and Space Agency (NASA) and the European Space Agency (ESA). NASA is in charge of the Saturn orbiter *Cassini*, and ESA is responsible for the *Huygens* descent probe to Titan's surface. Titan has been revealed as a close analog to primitive Earth, when life was just developing, and therefore it was targeted to be investigated even more intensely than its large parent Saturn.

Every person who has ever looked at Saturn through a telescope is fascinated by this mysterious object and intuitively wishes to study and to know more about this strange far away object. A feeling for its majesty and diversity is given by the beautiful NASA photomontage of a portion of the Saturn system in Fig. 1. The description of this strange world given by Sir William Herschel in 1805 cannot be improved upon.



Figure 1. Photomontage of a portion of the Saturn system. The picture simulates a spacecraft flying close to the large icy satellite *Dione* in the foreground. Saturn and its rings are in the background. Clockwise from the upper left, the satellites visible are Enceladus, Rhea, Dione, Tethys, Mimas and cloud-covered reddish Titan in the upper right (NASA picture). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

There is not perhaps another object in the heavens that presents us with such a variety of extraordinary phenomena as the planet Saturn: a magnificent globe, encompassed by a stupendous double ring: attended by seven satellites: ornamented with equatorial belts: compressed at the poles: turning upon its axis: mutually eclipsing its ring and satellites, and eclipsed by them: the most distant of the rings also turning upon its axis, and the same taking place with the farthest of the satellites: all the parts of the system of Saturn occasionally reflecting light to each other: the rings and moons illuminating the nights of the Saturnian: the globe and the satellites enlightening the dark parts of the rings: and the planet and rings throwing back the sun’s beams upon the moons, when they are deprived of them at the time of their conjunctions.

Saturn the Planet

Once Saturn was revealed in the telescope as its own world, like our Earth, it took a while to arrive at a reasonable physical model of the planet. Until about the mid-nineteenth century, astronomers subscribed to the opinion that Saturn could be inhabited by a race called Saturnians. It was deemed incredible “that so splendid a world as Saturn should be devoid of inhabitants.” However, misgivings were also voiced, and it was noted that the physical conditions on Saturn might not make the planet suitable for inhabitants of Earth. It was realized after some time that the density of Saturn and the other major planets was considerably lower than that of Earth. The notion that Saturn might be composed of material different from Earth was at first rejected, and its low density was explained by assuming that it is a hot distended gaseous globe that had not yet cooled down and solidified. Our modern knowledge of Saturn giving some basic selected parameters is summarized in Table 2.

Bulk Composition and Interior. Physically realistic models of the overall or bulk composition of Saturn and of its interior structure were first proposed in the 1920s and 30s and were improved and revised from 1950–1989. These models

Table 2. Selected Parameters of Saturn

Orbital	
Semimajor axis	9.539 AU (1427×10^6 km)
Period of revolution around Sun	29.46 years
Eccentricity of orbit	0.0557
Inclination of orbit to ecliptic	2.489°
Physical	
Rotation period (features at the equator)	10 ^h 14 ^m
Rotation period (magnetic field)	10 ^h 40.5 ^m
Inclination of equator to orbit	26.73°
Equatorial diameter	60,330 km
Polar diameter	54,950 km
Mass	95.15 m _⊕ = 5.686×10^{29} g
Bulk density	0.70
Escape velocity	37 km/s
Surface gravity	1040 cm/s ²
Equator centrifugal acceleration	176 cm/s ²

were aided considerably by advances in theoretical physics such as atomic and molecular theory and quantum mechanics and by improved laboratory data and accumulating observational constraints. The first models consisted of a solid interior overlain by a relatively thin atmosphere comprising about 20% of the planet's radius. To match the density of Saturn, it was supposed that the interior was made up of a thick layer of low-density ice that covered a small rocky core. The atmosphere was cold, and it was presumed to contain gases such as hydrogen, nitrogen, oxygen, helium, and perhaps methane whose freezing points are much lower than that of water. At the same time, it was recognized from the study of the abundance of the elements in the solar system and stars that hydrogen was by far the most abundant element. Its high abundance and low density, could possibly explain the bulk composition of the giant planets.

These early models overlooked two facts. The immense pressure in the interior of Saturn will increase the densities of even the lightest element such as hydrogen well above 1.0. Second, the insulating effect of Saturn's deep atmospheric blanket retains most of the primordial energy accumulated during accretion of the planet's mass. The interior of Saturn must therefore be quite hot. In fact, it is now known that Saturn exhibits heat flux from the interior equal to about 80% of the thermal energy it gathers from the Sun, or about $2000 \text{ ergs cm}^{-2} \text{ s}^{-1}$.

To obtain the best models of the interior of Saturn, the behavior of the elements and molecules, in particular, hydrogen and helium, under the high pressures and temperatures in the interior of Saturn, must be known. The pressures encountered exceed those that can be simulated in laboratory experiments so that recourse must be taken to theoretical quantum mechanical calculations. This has led to the interesting insight that when hydrogen is put under great pressure of the order of 3×10^6 atmospheres (3 Mbar), atoms or molecules lose their individuality and a high-density fluid of protons and electrons results. This state is quite conductive and has been given the label, metallic hydrogen.

There is no agreement on a unique model of Saturn's interior, but there is general consensus on its basic features. A typical model is shown in Fig. 2. The temperature at the 1-atmosphere (1-bar) level is about 140 K. Proceeding to the interior, the pressures and temperatures of Saturn's atmosphere keep rising, almost like diving into a limitless ocean, until the transition to metallic hydrogen (H^+) occurs about halfway into the planet at a pressure of about 3 Mbar (3×10^6 atmospheres). After the layer of metallic hydrogen gas, one possibly encounters a mantle of higher density ices, probably liquid, and a core of even higher density rock-forming elements, Si, Mg, O, and Fe. It is not certain whether the core is solid or liquid. A concentration of mass near the core is required by the higher order terms of Saturn's measured gravitational field. The mantle and core must contain about 20% of Saturn's mass. This makes the fraction of the elements that constitute the core and mantle considerably larger than for the solar composition for which it would be 2–3% of Saturn's mass. Thus, it is generally believed that Saturn's composition is not solar but is enhanced in rock-forming elements and/or ices by about a factor of 10. This enhancement must be accounted for in scenarios of Saturn's formation from the solar nebula.

Atmospheric Composition. Most of the direct compositional information that we have about Saturn comes from spectroscopy. Interestingly, Saturn had

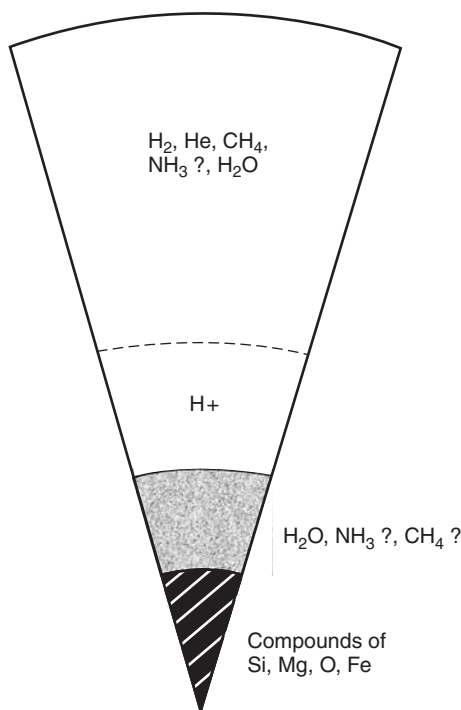


Figure 2. Model of the bulk composition and interior of Saturn after W.B. Hubbard and D.J. Stevenson. *Saturn*, T. Gehrels and M.S. Matthews (eds). University of Arizona Press, Tucson, 1984.

been observed visually with a spectroscope already in 1863 by the Italian astronomer, Father Secchi. He found dark unidentifiable absorption bands in the red part of the spectrum of the major planets. He felt that the atmospheres of these planets had not yet been “cleansed” and contained elements different from Earth’s atmosphere. These absorption bands were photographed in 1905 by E.C. Slipher of Lowell Observatory, but an identification was not made until 1932. They are caused by large amounts of gaseous methane in the atmosphere. The detection of molecular hydrogen, H_2 , although strongly believed present in even larger amounts, proved even more difficult. H_2 does not have a regular dipole spectrum in the infrared like most molecules, but exhibits only an exceedingly weak quadrupole spectrum. Lines from this spectrum were finally identified in 1963, and it was found that the amount of H_2 above the visible cloud tops of Saturn’s atmosphere is about 10 times the amount of the total Earth’s atmosphere. Further work in the 1970s and 1980s resulted in additional ground-based spectroscopic detection of the species PH_3 , C_2H_2 , C_2H_6 , and NH_3 .

A summary of the composition of Saturn’s atmosphere is given in Table 3. The major component is H_2 comprising about 92% of the atmosphere. Helium makes up most of the rest, but the exact fraction of He is difficult to determine. Its abundance cannot be measured by direct spectroscopic methods but must be inferred by its effect on other gases. It is estimated that the He abundance by

Table 3. Observed Composition of Saturn's Atmosphere

Species	Abundance	Comments
H ₂	~92%	Major atmospheric constituent
He	~8%	Inferred from secondary effects
H ₂ O	?	Presumed present but not detectable above 1 atm
CH ₄	0.2%	Enhanced by 2 to 3 over solar
NH ₃	0.03%	Frozen out above 1 atm
PH ₃	2 ppm	Enhanced by about 3 over solar
C ₂ H ₆	5.0 ppm	Abundance for upper atmosphere only where
C ₂ H ₂	~0.09 ppm	these molecules are produced by photodissociation

number of molecules (He/H₂) is about 6–10%. This number is quite uncertain, however, and needs further investigation and confirmation. If correct, it makes He about half as abundant as the 13% measured for the Sun.

It might be thought that the abundance of molecules is easier to determine for the extensive atmosphere of Saturn than for the modest atmosphere of its companion, Titan, but this is not the case. The level of the atmosphere probed depends on the wavelength of light. This makes it difficult to calculate the mixing ratios of gases detected in different spectral regions. In the deep atmosphere of Saturn, the most abundant molecule, after He, is thought to be water, but it has not been found because it is frozen out below altitudes to which spectroscopically detectable light reaches. Ammonia also freezes out but higher in the atmosphere, and so it has been observed. Its ratio varies with altitude and the exact level at which its presence has been detected is not certain. The abundance of methane is affected by the scattering properties of Saturn's atmosphere and is thus also not easily defined precisely. The best present determinations of the mixing ratios in Table 3 indicate that the species CH₄, NH₃ and PH₃ are enhanced about a factor of 2 to 3 over their abundance expected from solar composition. This lends observed compositional support to models of the interior of Saturn which point to a substantial enhancement of ice and rock-forming elements.

The variety and complexity of the molecules detected in Saturn cannot rival those of the atmosphere of Titan. The same Voyager infrared instrument that detected so many molecules in Titan could confirm only ethane (C₂H₆) and acetylene (C₂H₂) for Saturn. As explained in the section on Titan, the reason for this lies in the number of permissible photochemical reactions, which are considerably more limited under conditions of high H₂ abundance. The abundance of C₂H₆ and C₂H₂ given in Table 3 are for the stratosphere of Saturn. Their abundance drops by a factor of 10⁶ for the deep well-mixed Saturn atmosphere.

Visual Appearance and Colors. Early eighteenth and nineteenth century telescopic observers of Saturn focused much of their attention on the visual appearance of the planet. Investigators looked for the banded structure that is so prominent on Jupiter and tried to discover atmospheric features from which Saturn's rotation period could be determined. Yet, Saturn's cloud structure is

considerably different from that of Jupiter. While it shows some bands and zones, this structure on Saturn is very much muted and hard to see. The picture of Saturn in Fig. 3 is considerably enhanced and shows the structure typically displayed by Saturn. A faint slightly darker and yellowish band called the equatorial belt at latitudes of about 15° North or South can be seen and was described by many observers. It marks the limit of the bright equatorial zone. Sometimes another dark band, or possibly two, is seen at higher latitudes of about 30° . The polar regions above latitudes of $\sim 60^\circ$ are darker. The variable appearance of these features has been followed by observers from Cassini and Herschel to the modern Voyager spacecraft images.

Saturn is not round, but it exhibits considerable oblateness, the largest of any planet in our solar system. The equatorial radius is about 10% larger than the polar radius. This is caused by Saturn's rapid rotation. The centrifugal acceleration at the equator is 176 cm sec^{-2} , which is roughly 17% of its gravitational attraction. The resulting attractive force at the equator is thus considerably less than at the poles allowing the material at the equator to spread to a much larger radius.

One would very much like to have an explanation for the different colors of the zones and belts that have been admired by observers during the last few hundred years. Yet this has been achieved neither for Jupiter nor for Saturn. The subdued tones, it is believed, are caused by coloring agents or chromophores that are present in such minute amounts that they have not yet been identified. The suggestion has been made that the colors are caused by dilute allotropes of sulfur or phosphorous which can have a yellow or red or even black appearance.

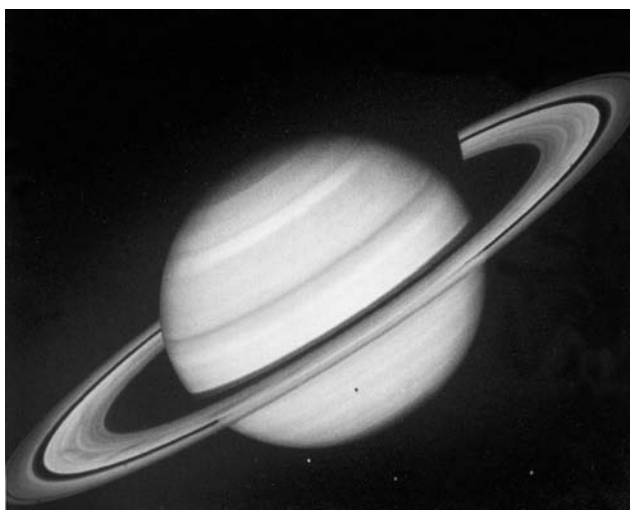


Figure 3. Image of Saturn and its rings floating mysteriously in space taken by the Voyager 2 spacecraft. The sun comes from below so that the ball of Saturn casts a shadow on the rings, and the inner Crepe Ring in turn forms a shadow on Saturn. Also visible are three icy satellites Tethys, Dione and Rhea in the foreground. The shadow of the leftmost satellite Tethys is projected onto the atmospheric cloud deck of Saturn (NASA picture). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Triclinic solid phosphorous (P_4) is red, and allotropes of phosphorous are likely because PH_3 has been detected and can be photodissociated high in the atmosphere. Sulfur-bearing molecules such as H_2S or NH_4HS are frozen out lower in Saturn's atmosphere and are less likely to be convected to great heights.

Atmospheric Structure, Clouds. A temperature profile for the observable part of Saturn's atmosphere is given in Fig. 4. The temperature profile in Fig. 4 is determined from radio occultation and infrared experiments onboard Voyager. The minimum temperature of about 85 K occurs at the tropopause at a pressure of about 0.10 atm. (This is similar to that level on Earth.) Above that level, in the stratosphere, the temperature rises slightly before becoming isothermal. Below that level, the temperature increases at a rate of about 0.9 K/km. This rate is called the adiabatic lapse rate, and it is expected if the atmosphere is strongly convective. At a level of about 1 atm, which is roughly the limit of our remote sensing observability of Saturn, the temperature has risen to 140 K. Near this level, the cloud density becomes so thick (optical depths 3–5), that even

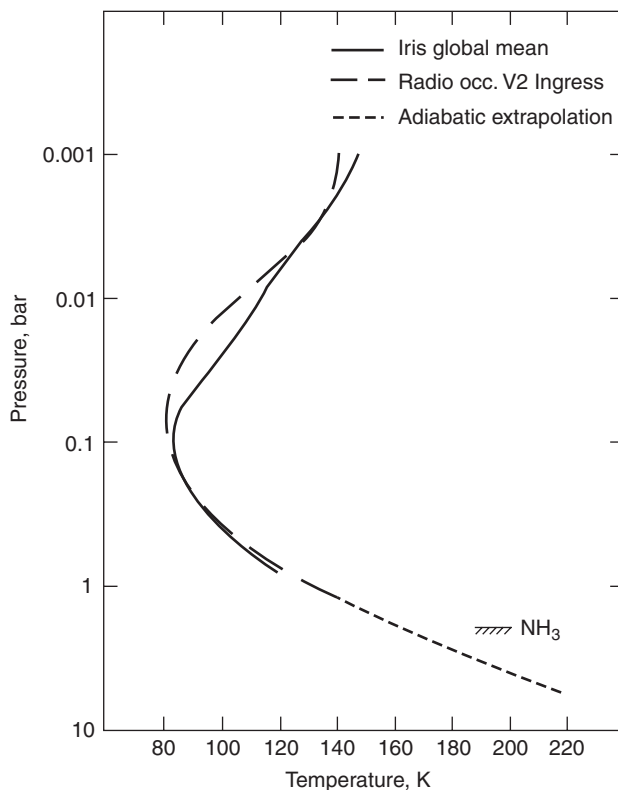


Figure 4. Temperature profile of the atmosphere of Saturn that is accessible to remote sensing. The temperature profiles are from measurements by the Infrared Imaging Spectrometer (IRIS) on the Voyager spacecrafts, the Voyager Radio occultation measurements, and an adiabatic lapse rate model for the deeper atmosphere. Graph is from A.P. Ingersoll et al. *Saturn*, T. Gehrels and M.S. Matthews (eds). University of Arizona Press, Tucson, 1984.

long-wave infrared radiation cannot penetrate below. It is presumed that this cloud deck is made of ammonia crystals, and there is some evidence for solid NH_3 absorption in infrared spectra, but this is not definite. The temperature at the 1-atmosphere level is still too cold to allow an appreciable amount of gaseous NH_3 to exist in the troposphere.

The atmosphere above the thick cloud deck is not clear but contains a haze of NH_3 crystals extending to the tropopause. This region contains different dilute concentrations of chromophores that create the different colors among the belts and the zones. It is known that the zones are regions of atmospheric upwelling while the dusker belts are regions where cold stratospheric air descends. The colder troposphere of Saturn, compared to that of Jupiter probably results in a paucity of upwelling chromophoric parent molecules, such as PH_3 , leading to a lower concentration of coloring agents and a subdued contrast between the belts and zones.

Rotation Rate and Circulation. The determination of Saturn's rotation rate by using visible atmospheric features was quite difficult, compared to Jupiter, because well-defined features in Saturn's atmosphere are rare. William Herschel in 1794 was the first to measure a rotation rate, and his value of $10^{\text{h}}16^{\text{m}}$ is close to the present value of $10^{\text{h}}14^{\text{m}}$ for features near the equator. After Herschel, well-defined spots on Saturn were not observed until 1876, and this was followed by a series of spots observed from 1891–1903. It was realized that Saturn has a very pronounced differential rotation rate; spots at latitudes $\sim 35^\circ\text{N}$ traverse Saturn considerably more slowly in a time of about $10^{\text{h}}38^{\text{m}}$. The most accurate rotation rate today comes from measurements of the rotation of Saturn's magnetic field which is $10^{\text{h}}40.5^{\text{m}}$. It is felt that this represents the rotation period deep inside Saturn's conductive metallic hydrogen mantle.

The rotation rate of surface spots can be used to measure the upper atmospheric currents or the weather on Saturn. It is found that the wind speeds in Saturn's upper equatorial zone are about 500 m s^{-1} , which is 1100 miles/hour! The same phenomenon is found on Jupiter, but that planet's equatorial wind speed is about 120 m s^{-1} , while jet streams on Earth have typical speeds of perhaps 100 miles/hour. Besides having an equatorial wind speed four times higher, Saturn's equatorial zone is also about twice the width of Jupiter's. It is not yet clear what causes the large difference between the planets. It could be attributed to a combination of Saturn's lower surface gravity and its lower internal heat flux. Another interesting question, not presently answered, is the depth of the zones of differential rotation. It is conjectured that they may persist to 10,000 km or more into the interior of the planet. The "weather pattern" on Saturn may thus be of an entirely different nature from that observed on Earth.

The Rings of Saturn

Galileo was the first to describe the Saturn system, but he was confused by its appearance and could not come up with a correct explanation. The correct interpretation for the rings of Saturn and their changing appearance was first given by Christiaan Huygens in 1655: "*Annulo circigitur tenui, plano nusquam cohaerente...*" "a thin ring surrounding Saturn nowhere touching the planet". His

discovery of Saturn's satellite, Titan, and its 16-day orbit around Saturn helped him realize that the line of ansae (i.e., the ring plane) was tilted about 27° with respect to Saturn's orbital plane. With the ring plane fixed in space, to a viewer on Earth, the rings would change from perfectly edge-on to completely open, as Saturn travels around its orbit. This resolved Galileo's mystery when 2 years after he viewed the rings, they mysteriously disappeared leaving him baffled and confounded.

When viewed through the telescope, the rings appear to the eye as a solid sheet separated by the Cassini division. Early investigators thus believed the rings to be made of two solid annuli. But how were these annuli held in place around Saturn and did they have a rotational period? Their uniform appearance foiled every realistic attempt to determine a rotational period. The French mathematician Laplace presented calculations in 1785 which showed that the rings must be made of many individual narrow rings and that the rings must rotate about the planet in Keplerian orbits, that is, each narrow ring must circle the planet with a rotation period given by Kepler's third law. This theory was expanded in 1857, by James Clerk Maxwell who showed that the rings must be made of individual particles or satellites. These objects were so densely packed that to the eye they gave the appearance of a solid continuous sheet. The fact that the rings rotate at their respective Keplerian velocities was proven in 1895 by the American astronomer James E. Keeler using the new science of spectroscopy. As the ring particles rotate about Saturn, those on one side are coming toward us, whereas those on the other side are moving away from us. By the Doppler principle, this produces a wavelength shift to the blue and to the red, respectively. The velocities of the particles on the outside edge of the ring are slower than those on the inside edge, resulting in tilted spectral lines. This Doppler shift tilt was large enough that it could be captured on the photographic plates of that time, giving observational proof of the rotation of the rings and their composition of individual particles.

The major ring system, its size, and nomenclature are given in Table 4. The reason for the somewhat confusing sequence of names is historical. The two bright rings that can most easily be seen in the telescope are called A and B (starting with the ring farthest from Saturn). The A and B rings are separated by the Cassini division discovered by the Italian/French astronomer, Jean Dominique Cassini, in 1675. A fainter ring, called the C ring or Crepe ring, inside ring B was first described by the American astronomer, W.C. Bond (1850). Inside the C ring, astronomers using ground-based telescopes thought they saw another faint ring called D. This ring could not be detected by the Pioneer 11 spacecraft. However, the Voyager spacecraft did find an exceedingly faint ring system near that position. Even though the ring was so faint that it could not have been detected from the ground, its assignment as D ring was kept. A faint diffuse ring outside of the A ring was discerned using ground-based telescopes in 1967 and was called the E ring. It was verified by the Voyager spacecraft. The detection of a very narrow ring 50 km wide outside the A ring, which was called the F ring, was one of the interesting discoveries by the Pioneer 11 spacecraft. The Voyager spacecraft also imaged a very diffuse ring between the F and E ring which was labeled G. It was first detected indirectly by the Pioneer 11 spacecraft because of its influence on charged particles.

Table 4. The Rings of Saturn: Nomenclature, Size, Features

	Distance from Saturn center		Features
	km	R_s	
Saturn boundary	60,330	1.00	Top of visible atmosphere; equatorial radius
D ring	67,000–73,200	1.11–1.21	Quite tenuous, discovered by Voyager
C ring	74,500–92,200	1.23–1.53	Crepe ring, Bond(1850)
B Ring	92,200–117,500	1.53–1.95	Biggest and most extensive ring; easily seen with telescope
Cassini division (middle)	119,000	1.97	Discovered by Cassini in 1675
A ring	121,000–136,200	2.03–2.27	Outermost ring seen with telescope
Encke gap (middle)	133,600	2.21	Discovered by Encke; quite sharp, ~328 km wide
F ring (middle)	140,400	2.33	Discovered by Pioneer; quite narrow, ~50 km
G ring (middle)	170,000	2.82	Discovered by Pioneer
E ring	181,000–483,000	3–8	

The illusory telescopic view of the rings as a solid sheet has already been discussed. During the centuries, many different divisions were spotted and described by observers, but except for the obvious Cassini division and the fleeting Encke gap, no consensus by any observers could be established. Not many people, however, were prepared for the astonishing images sent back by Voyager which showed thousands of individual ringlets. A sample image is provided in Fig. 5. Even Voyager’s best resolution of about 10 km could not reveal the wealth of detail present. A stellar occultation with the Voyager photopolarimeter instrument indicated that there must be tens of thousands of ringlets. Additionally, quite unsuspected anomalous features such as several noncircular ringlets that have notable eccentricities and several “braided” ringlets were observed.

A very puzzling transient phenomenon of radial spokes in the B ring was observed by Voyager. Those can be seen quite clearly in Fig. 5. The spokes generally rotate around Saturn with the Keplerian velocities of the particles and thus lent themselves to several Voyager imaging time sequences that were assembled into a captivating movie. Most likely, the spokes are made of micron-sized particles slightly above the plane of the rings, whose light scattering properties makes them visible, and their formation is connected with Saturn’s magnetic field.

Despite the very large planar extent of the rings, about 61,700 km for the A, B, and C rings, the rings are quite thin. Ground-based observations during Saturn’s ring plane crossing yielded an upper limit of about 2 km. Voyager could not measure the exact thickness but determined an improved upper limit of 100–200 m. To get a feeling for the extreme thinness of the ring, one should

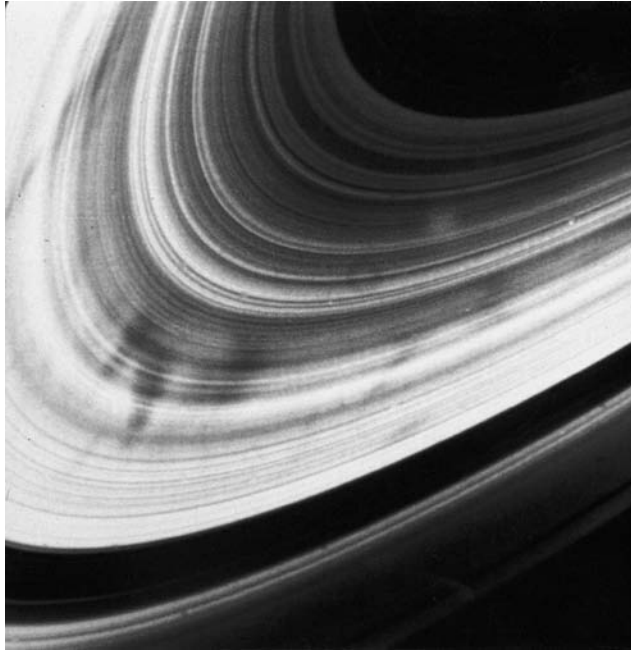


Figure 5. Close-up Voyager 2 picture of Saturn's rings showing its structure as a multitude of ringlets and the mysterious dark radial features or 'spokes' (NASA picture). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

realize that this is equivalent to an annular sheet of paper that has an outer radius of 400 feet and an inner radius of 220 feet! The rings of Saturn are therefore considerably less than *paperthin*. The features of the rings that we would like to know about are the thickness of the rings, the cause for the ring divisions, the compositions of the rings, the particle size distribution of the rings, and the total amount of material in the rings. These subjects are discussed in turn later, and though we have many partial answers, more investigation is required.

A very interesting additional phenomenon found by Voyager spacecraft was the dynamic nature of the rings. Sequences of images showed that the ringlet structure is not fixed but appears to change from orbit to orbit. These small rearrangements, like waves on an ocean, occur even though the ring system as a whole is stable for millions and probably billions of years. There is no complete explanation for the ordering and arrangement of the ringlets. Density waves appear to account for many features, particularly in the A ring. Gravitational interactions with Saturn's satellites, called resonances, account for more of the features seen. However, the resonances do not offer a ready explanation for many of the prominent major gaps, even the Cassini division. The mystery of the presence of the Encke gap was solved in 1991 by going back through the large Voyager data archive and discovering Saturn's eighteenth satellite, Pan, in this gap. Perhaps, many of the other minor ring divisions are kept clear by smaller yet still substantially sized objects several km in diameter.

The narrow F ring has a straightforward explanation. Voyager found two satellites, one on the inner side of the F ring (S16) and one on the outer side of the F ring (S15). Like sheep dogs herding a flock of sheep, these two satellites keep the material from spreading, and therefore they are called “shepherding satellites.”

Despite the sophistication of the Voyager spacecraft and its close approach to Saturn, its instruments could not determine the composition of Saturn’s rings. This was first done in 1970, from ground-based infrared spectra. It was found that the particles in Saturn’s rings were predominantly water ice. It is believed that some of the ice may have a core of silicate dust or that pure silicate dust particles are also present. This is particularly so for the C ring which is considerably darker than the A or B ring. The temperature of Saturn’s rings, as measured by their infrared emission, is around 90 K, which is cold enough that no significant amount of water molecules evaporate during a span of a billion years.

The size of the particles in Saturn’s rings must clearly be less than the upper bound of its thickness. The size range of the particles is estimated from the scattering properties of the rings at wavelengths from the visible to the microwave region. Particularly helpful were radio occultations by the rings of the 3.6-cm (X-band) and 13-cm (S-band) radio transmitters on Voyager. Combining these data leads to the conclusion that the particle sizes range from micron-sized grains to meter-sized boulders, and the upper limit is of the order of 10 m in diameter. The particle size distribution (cumulative number of particles versus size) appears to fall off with particle radius a as a^{-3} .

The total mass of Saturn’s rings is also a quantity that must be estimated by indirect means. Pioneer 11 passed within 4000 km of the A ring but experienced no gravitational perturbation by the ring. This gives an upper limit of $1.7 \times 10^{-6} M_s$. A variety of indirect measurements converges on a mass of about $5 \times 10^{-8} M_s$, or 2.8×10^{22} g. This is about 1/2600 the mass of Earth’s Moon, or assuming a density of 1.0, is equivalent to a spherical satellite of 380 km in diameter, about the size of Mimas.

Satellites of Saturn

Presently, 30 satellites that circle Saturn are known. This gives Saturn the largest number of satellites in our solar system. Following Huygens’ discovery of Titan in 1655, four more satellites were soon found by Cassini: Iapetus (1671), Rhea (1672), and Dione and Tethys (1684). Since that time, the number of Saturn’s satellites has steadily increased. By 1900, nine satellites were known (S1–S9), and these are often called the nine classical satellites of Saturn. Preparations for the Voyager spacecraft spurred renewed interest in searching for Saturn companions, so that during the two ring plane crossing periods, 1966 and 1980, five more satellites (S10, S11, S12, S13 and S14) were discovered using Earth-based telescopes. These discoveries were aided considerably by the invention and introduction of large-area silicon array detectors called CCDs (charge-coupled devices). The Voyager spacecraft itself found an additional four satellites (S15, S16, S17, and S18) so that the number of known satellites rose to 18. More

recently, astronomers from Cornell University obtained a number of deep CCD images of the surroundings of Saturn and found 12 more satellites that have been given the temporary designation S/2000 S1 to S/2000 S12 for the year of their first observation, 2000. As an aside, we note that this is much greater than the number of six to eight satellites, which Kepler (1571–1630) conjectured would be found around Saturn. It is unlikely that any sizable satellite of the order of hundreds of km diameter remains undiscovered, but it is quite likely that there are more 10-km sized objects. A list of Saturn's satellites and some of their basic parameters are provided in Table 5. In this section, we discuss the properties of all of the satellites except Titan, which is so unusual that it deserves a special section.

As for the names of the satellites, we note that it was John Herschel, the son of Sir William Herschel, the discoverer of Uranus, who bestowed the present names on the Saturnian satellites. Following the Greek legend of Saturn, they are named after Titans or giants, either brother or sister siblings of Saturn. Before Herschel, Titan was generally known as the "Huygenian satellite," and Cassini tried to name the four satellites he discovered after his sovereign Louis XIV. The satellites also have numbers (S1–S18) where the S stands for Saturn. The first nine are neatly ordered by increasing distance from Saturn. Thereafter, their sequence follows the order of their discovery to avoid constant renumbering when new ones are found.

Astronomers in the nineteenth century concentrated on understanding the celestial mechanics of the satellite system by getting good positional measurements from which accurate orbital parameters could be derived. They found that the mechanics of Saturn's satellite system was more intricate and complex than might be expected. John Herschel was the first to note that, "the time of revolution of Mimas is very close to half that of Tethys, and that of Enceladus approximates half that of Dione." Since that time, these so-called commensurabilities or resonances have been studied intensely with increasing sophistication by investigators in celestial mechanics. If we use the following abbreviations for the mean daily motion of a satellite going around Saturn in deg/day: Mimas-Mi, Enceladus-En, Tethys-Th, Dione-Di, Titan-Ti and Hyperion-Hy, the relationships below can be written:

$$Mi = 2Th + 0.84 \text{ deg/day},$$

$$En = 2Di - 0.41 \text{ deg/day},$$

$$3Ti = 4Hy + 0.060 \text{ deg/day}.$$

Thus, for example, an exact orbital 2:1 resonance between the orbital period of Enceladus and Dione is off by only one orbit out of 640. More sophisticated orbital mechanics can explain this small difference by the slow precession of the nodes (the line of intersection of the plane of the satellite orbit with the plane of Saturn's orbit).

Jumping ahead for a moment to modern investigations, the Pioneer encounter, the Voyager spacecraft images, and supporting ground-based observations

Table 5. **Satellites of Saturn**

	Name	Dist. from Saturn, R_s	Orbital period, hrs	Size, km	Density, g/cm^3	Comments
S18	Pan	2.21	13.8	20	—	In Encke's gap
S17	Atlas	2.28	14.4	$40 \times ? \times 20$	—	A-ring shepherd
S16	Prometheus	2.31	14.7	$140 \times 100 \times 74$	—	F-ring inside shepherd
S15	Pandora	2.35	15.1	$110 \times 90 \times 66$	—	F-ring outside shepherd
S10	Janus	2.51	16.7	$220 \times 190 \times 160$	—	Co-orbiting and
S11	Epimetheus	2.51	16.7	$140 \times 115 \times 100$	—	Oscillating satellites
S1	Mimas	3.08	22.6	394	1.4	Densely cratered; Herschel 125 km diam.
S2	Enceladus	3.95	32.9	502	1.2	Large smooth lightly cratered surface
S3	Tethys	4.88	45.3	1060	1.2	Densely cratered; Odysseus 400 km diam.
S13	Telesto	4.88	45.3	$30 \times 20 \times 16$	—	Tethys leading Lagrangian satellite
S14	Calypso	4.88	45.3	$24 \times 22 \times 22$	—	Tethys trailing Lagrangian satellite
S4	Dione	6.26	65.7	1120	1.4	Large surface brightness variations
S12	Helene	6.26	65.7	$34 \times 32 \times 30$	—	Dione leading Lagrangian satellite
S5	Rhea	8.74	108	1530	1.2	Moderately cratered; (2 populations?)
S6	Titan	20.3	383	5150	1.88	Extensive atmosphere
S7	Hyperion	24.6	511	$410 \times 260 \times 220$	—	Irregular shape
S8	Iapetus	59.0	1904	1440	1.2	Large albedo difference in E, W faces
S9	Phoebe	215	13211	220	—	Retrograde orbit; dark surface; captured?
S/2000	—	—	—	—	—	Twelve additional satellites discovered
S1–S12						using ground based CCD observations;
						no permanent name assignments yet;
						sparse physical information

revealed even more surprises in the mechanics of the Saturn satellite system. Two of the satellites, Janus (S10) and Epimetheus (S11), actually share the same orbit. It was found that the satellites slowly drift toward each other in the bit, yet a collision is avoided because their mutual gravitational interaction, shifts their orbits slightly, when they get close to each other, and they drift apart again to meet on the other side of the orbit, whence the whole sequence repeats. The large satellite, Tethys, was also found, to have two coorbital objects, Telesto (S13) and Calypso (S14). These, however, remain a constant distance from Tethys since they are located at the stable Lagrangian points 60° ahead and 60° following Tethys. Dione, it was also found, has a satellite, Helene (S12), at its leading Lagrangian point. A careful search by Voyager at the following Lagrangian point of Dione failed to show any object.

Four satellites (S15–S18) all have strong connections with and influence on Saturn's ring system. The Voyager spacecraft found that the narrow F ring was restrained from spreading by two shepherding satellites, one on the outside of the ring, Pandora (S15), and one just inside the ring, Prometheus (S16). A small satellite, Atlas (S17), was found just outside the A ring and, it is believed, keeps this ring from spreading outward, thus accounting for its sharp boundary. Finally, careful inspection of the 30,000 Voyager images of Saturn and its rings in 1991 revealed an additional small satellite, Pan (S18), in the Encke gap. The satellite keeps this gap clear and provides a reason for Encke's gap at this particular location in the rings.

Limited to the naked eye, classical telescope observers found it difficult to determine physical properties of the satellites, because even with reasonably sized telescopes, the satellites appeared as mere faint points of light. It was realized, since it was discovered by Cassini, that Iapetus is much brighter on the western side of Saturn (its trailing side as it orbits Saturn) compared to the eastern side. As Cassini stated, "This satellite disappears regularly for about one-half of its revolution when it is to the east of Saturn." Modern photoelectric measurements have determined that this variation is 1.7 magnitudes, or a factor of 4.8! This light curve, invariable for centuries, is considered proof that Iapetus keeps the same face toward Saturn as it orbits that planet, just as our moon shows the same face to Earth. Iapetus' rotational period is synchronously locked with its rotational period around Saturn.

Variations in visual orbital brightness were sought for the other satellites but yielded conflicting results. In the 1970s ground-based electronic photometry revealed that the satellites Rhea, Dione, and Tethys differ from Iapetus and are all brighter on the leading side, respectively, by 0.20, 0.40 and 0.20 magnitudes. On the other hand, Mimas is brighter on the trailing side by about 0.2 magnitudes. Most of Saturn's satellites for which rotation periods have been determined, namely S1, S2, S3, S4, S5, S8, S10, and S11, are synchronously locked to Saturn. The rotation period of the outermost satellite, Phoebe, is 9.4 hours, and Hyperion does not have a stable rotation period but tumbles chaotically.

Attempts were made to determine satellite sizes during periods of superior seeing, and sizes were estimated from their brightness by assuming a reasonable reflectivity. The theory of orbital resonance, discussed earlier, allowed determining

the masses of some satellites. Combining the mass and size made estimates of the satellites' densities possible. This led to the interesting realization in the 1920s that satellite densities were probably best described by a value airy close to one which is similar to the density of water ice. This density is considerably different from that of rocky material whose values are 2.5–3.0. The icy composition of Saturn's satellites has been borne out by modern measurements. The best present densities that combine data from the Voyager spacecraft and ground-based observations are given in Table 5. The listed densities are all slightly higher than 1.0, but it must be remembered that the density of ice increases as pressure increases in the interior of the satellites and somerocky material is probably mixed with the ice, both on the surface and in the interior.

Electronic photometry in the twentieth century was used to determine accurate values for the brightness, or visual magnitude, of the satellites, as well as their broadband colors in the violet, blue, visible, and red. These measurements showed that the surface reflectivity, or albedo, of the satellites was quite high, typically of the order of 50–60%. The highest value measured was 90% for Enceladus. A material on Earth that has this high reflectivity would be brilliantly and blindingly white. Improved measurements in the thermal emission regime showed that the surface temperatures were quite low, of the order of 90 K, even in the warmest regions near the equator. Probably the most significant ground-based results were provided by infrared spectroscopy of Iapetus, Rhea, Dione, and Tethys in 1976. These data unmistakably showed the unique signature of water ice. Combining this direct proof of the existence of water ice with the low density of the satellites and their low surface temperatures leads to the conclusion that the large spherical satellites of Saturn are composed of water ice. They are thus labeled Saturn's icy satellites.

To make further progress in our understanding of the satellites, close-up views were needed and these became available the beginning of the Space Age. Following the exploratory Pioneer missions, the two Voyager spacecraft imaged all fourteen satellites known at that time, plus four additional ones that were discovered.

Having roughly the same density and being of comparable size, it might be expected that the satellites all had a similar origin and thus closely similar appearances, but this was not found to be the case. Each satellite turned out to have its unique individual characteristics, making them distinctly different from each other. The six major icy satellites, Mimas, Enceladus, Tethys, Dione, Rhea, and Iapetus, are shown together with the much smaller bodies, Hyperion and Phoebe, in Fig. 6. All of them have sufficient self-gravity to make them spherical, and all of them possess cratered surfaces to various degrees. At the low surface temperatures of these satellites, water ice is as rigid as ordinary rock and can easily retain a cratering history of a billion years or so. Curiously enough, and perhaps not by accident, the six large icy satellites of Saturn show a gradation in size as one moves outward from Saturn; Mimas and Enceladus are about 450 km, Tethys and Dione about 1100 km, and Rhea and Iapetus about 1500 km in diameter. Although the satellites in Fig. 6 may look similar to our Moon, we tend to forget that these are icy bodies that have high reflectivities and would be brilliantly white if

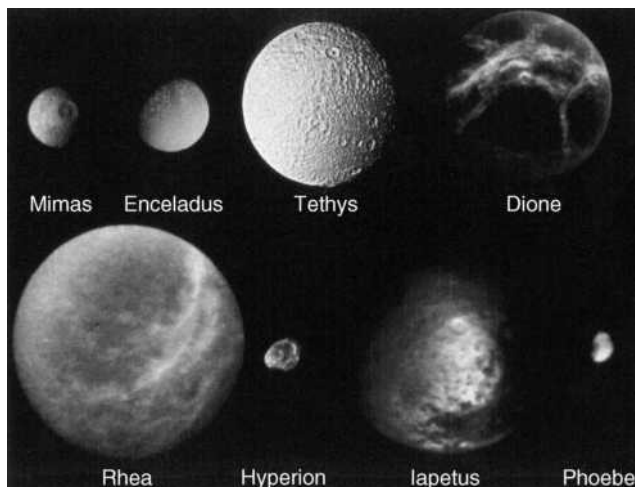


Figure 6. Composite of the icy satellites of Saturn. Roughly 450 km diameter satellites *Mimas* and *Enceladus* are at the top left, then come the roughly 1100-km diameter satellites *Tethys* and *Dione*. The bottom row shows the two larger 1500-km diameter satellites *Rhea* and *Iapetus*, as well as the two much smaller and irregularly shaped objects *Hyperion* (about 300 km) and *Phoebe* (about 220 km). Ground-based observations have established the presence of water ice on the surfaces of all of these objects except *Phoebe*. The brightness of *Phoebe* has been enhanced considerably; it is actually a very dark object whose reflectivity is close to that of the dark side of *Iapetus* (NASA composite). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

one could stand on their surface; our Moon has a very low albedo of 0.06 and is quite dark.

Mimas. This innermost satellite of the large icy satellites is quite heavily cratered, and its surface shows little evidence of interior activity. Its appearance is dominated by one large crater, “Herschel,” on its leading surface. The crater is 130 km in diameter and thus covers about one-third of the satellite’s diameter.

Enceladus. This satellite is almost the same size as Mimas, but it has considerably different surface features. There are sizable areas on this satellite that are moderately cratered (though not as densely as Mimas); its most noteworthy features are large uncratered smooth plains. These plains are traversed by very noticeable ridges and grooves. It appears that large areas of this satellite were flooded by liquid water from the interior, which formed the smooth plains and obliterated any previous cratering history. The extensive icy plains of this satellite give it very high reflectivity. It is not yet clear what made this satellite so different in appearance and geologic history from its twin, Mimas.

Tethys. This satellite is densely cratered across its whole surface, similar to Mimas. It presents almost no evidence of geologic surface activity. Interestingly enough, similar to Mimas, Tethys has one very large 400-km diameter impact crater “Odysseus.” Although the extent of this crater is 40% of the satellite’s diameter, it does not dominate the surface as Herschel does for Mimas, because Odysseus has lost its sharp boundaries and deep interior and has “relaxed” considerably. An additional feature of Tethys is a 2000-km long

valley called Ihaca Chasma, that stretches three-quarters of the way around its circumference.

Dione. Dione has the highest density of any of the icy satellites and thus probably the highest component of rocky material. This is corroborated by its relatively low albedo and the weaker signature of water ice on its surface. Its overall crater density is relatively low, and there is no significant difference in crater density between the leading and trailing sides. There is, however, a distinct difference in visual appearance between these two sides. The trailing hemisphere is darker by 0.4 magnitudes (a factor of 1.5) and shows strong albedo markings of global dark terrain traversed by half a dozen broad white bands, some extending across the whole diameter. These white bands have a reflectivity of 0.7 versus the general background reflectivity of the dark material of 0.2. It is likely that the bright bands were created by a global system of faults that allowed fresh clear ice to cover the surface. Dione has several terrains that have different crater densities and a considerable level of geologic activity and resurfacing.

Rhea. Rhea is the largest of the icy satellites. Similar to Dione, it has a trailing hemisphere that is darker, though only by 0.2 magnitudes (a factor of 1.2), and like Dione, it has several bright bands extending across the otherwise darker trailing hemisphere. Rhea is fairly heavily cratered, and there appears to be no difference in cratering activity between the leading and trailing side. There are indications that it is an older terrain that retained craters larger than 30 km in diameter and a younger resurfaced terrain, but there is no consensus on this point.

Iapetus. The huge albedo difference between Iapetus' leading and trailing sides has already been mentioned; the leading side has a reflectivity of 4%, as dark as coal dust, whereas the trailing side is quite bright and has a reflectivity of 30–40%. It might be thought that a few good images of the satellite would have resolved this question, but this was not the case. The Voyager spacecraft did not get close enough to Iapetus for good high-resolution pictures, and the dark areas were so dark that the cameras could not pick up any detail! The bright side of Iapetus is quite heavily cratered. Because its density is 1.1, Iapetus clearly must be composed of water ice. Thus, the reason for the bright and dark hemispheres of Iapetus still remains a mystery.

Other Satellites. The two other moderate sized satellites, Hyperion and Phoebe, are quite different from each other. Spectroscopy indicates that Hyperion belongs to the class of icy satellites. Its size of 300 km argues for a spherical shape, yet it is quite irregular and probably is a leftover fragment of a past collision. Phoebe on the other hand, although slightly smaller, appears to be spherical. It is in a retrograde orbit that makes it an irregular satellite, and it has a dark surface and a visual reflectivity that is better matched by rocks.

A composite picture of the small irregular shaped satellites is shown in Fig. 7. The exceedingly interesting dynamic properties of S10–S18 have already been discussed. The Encke gap satellite S18 shows up only as a bright dot in the Voyager images and so allowed no shape inferences. The Voyager imaging was good enough that the remaining small satellites (S10–S17) all showed distinct irregular shapes. They appear to be fragments of collisions, very likely captured.



Figure 7. A selection of eight irregularly shaped satellites of Saturn. The leftmost satellite is *Atlas* (S17), the A-ring shepherd. The next pair (top to bottom) are the F-ring shepherds *Pandora* (S15) and *Prometheus* (S16); then follow the two coorbiting satellites *Janus* (S10) and *Epimetheus* (S11); then come the two Tethys *Lagrangian* satellites *Calypso* (S14) and *Telesto* (S13), and finally at the right is the Dione leading *Lagrangian* satellite *Helene* (S12) (NASA photomontage). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Some of the larger ones have a discernible number of craters on their surfaces. Most of them appear to be of icy compositions. Not much is known about the recently discovered satellites S/2000 S1–S12, except that they are relatively small, from 10–50 km in diameter, and are irregular satellites.

Titan

The largest moon of Saturn, Titan, has turned out to be one of the most interesting satellites in the solar system. With a diameter is 5150 km, it is one of the largest solid objects in the solar system, only slightly smaller than Ganymede of Jupiter, but considerably bigger than the planet Mercury, and easily dwarfs our most distant member, Pluto. When Titan was discovered by Huygens in 1655, no one could have foreseen that it held so many surprises. For centuries, people thought of it as just another moon whose position, orbit, brightness, transits across Saturn, etc. needed to be duly observed and recorded. It was not until modern instrumentation was employed that Titan revealed its unusual nature. Selected physical parameters for Titan are listed in Table 6.

Using infrared spectroscopy, G.P. Kuiper in 1943 detected deep methane absorptions, similar in strength to those of other major planets. This was quite astonishing for a satellite. Further observations, in the 1970s, of stronger methane bands farther in the infrared, combined with careful laboratory comparison spectra, led to the conclusion that the amount of methane on Titan was of the

Table 6. **Selected Physical Parameters for Titan**

Diameter	5150 km
Mass	$1.35 \times 10^{26} \text{ g} = 0.022 \text{ m}_{\oplus}$
Density	1.88
Surface gravity	135 cm s^{-2}
Temperature	
Surface	94 K
Effective	86 K
Tropopause(42 km, 128 mb)	71 K
Pressure surface	1.50 atm
Ratio rock/ice, by mass	52/48
Hydrogen (H) loss rate	$5 \times 10^{27} \text{ atoms/s}$

order of 20% of the total Earth’s atmosphere; a very substantial atmosphere twenty times larger than the atmosphere on Mars. More advances in infrared instrumentation in the 10 to 20 μm region revealed the presence of ethane (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2).

Ultraviolet data showed that the atmosphere of Titan must be quite hazy and filled with aerosols (small smog-like particles) that scatter the incoming radiation. There were suggestions that Titan’s atmosphere might be more substantial requiring the additional presence of spectroscopically nondetectable gases such as N_2 , Ne and Ar. From the limited amount of ground-based data, it was not possible to separate the effects of scattering from those of a denser atmosphere and a determination of the correct amount of CH_4 in a vertical column of Saturn’s atmosphere remains elusive. It was clear, however, that Titan’s atmosphere exhibited a most interesting character.

The Voyager 1 spacecraft was therefore, targeted to pass very close to Titan, within 4000 km of its cloud tops. The data acquired were both disappointing and fascinating. Despite the excellent resolution of Voyager’s cameras, the hazy atmosphere prevented any glimpse of the surface, a surface which many investigators believe could be one of the most unusual and spectacular in our solar system. The hazy atmosphere was so pervasive and uniform that it took special image processing techniques to bring out a very faint banded structure and a small brightness difference between the Southern and Northern Hemispheres. Such an image is shown in Fig. 8. Pictures of the limb of the satellite proved more successful and showed several haze layers; the highest was 700 km above the surface. An image of this haze is presented in Fig. 9. (On Earth, the clouds and haze rarely reach above about 12 km.)

On the other hand, the atmospheric data returned were exceedingly interesting and fascinating. Three instruments provided the most important data. The ultraviolet spectra showed definite evidence of N_2 , the radio occultation experiment provided the atmospheric temperature profile and the surface pressure, and the infrared instrument discovered a whole host of new organic molecules, as well as temperature data and information on the atmospheric structure. In fact, one might say that the capabilities of the infrared instruments were as well matched to the investigation of Titan’s atmosphere, as might

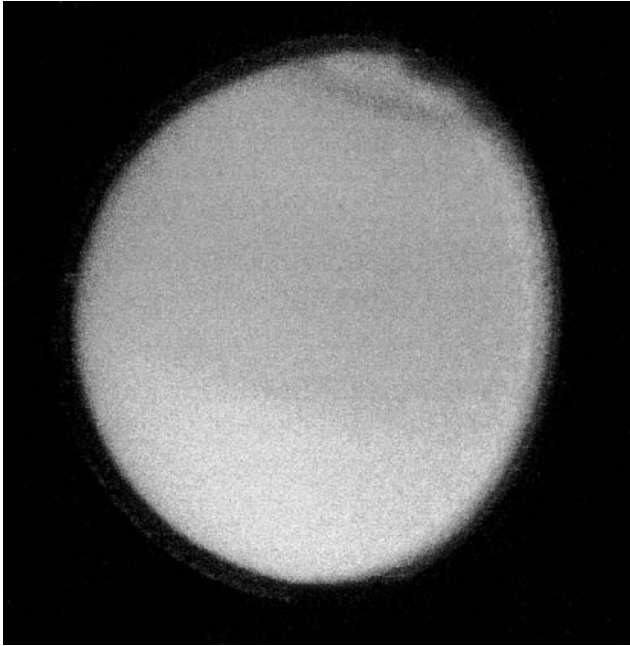


Figure 8. Voyager 2 image of *Titan* showing the thick, reddish cloud layer that obscures any surface features. A faint dark band is visible near the North Pole and the Southern Hemisphere is slightly higher (NASA picture). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

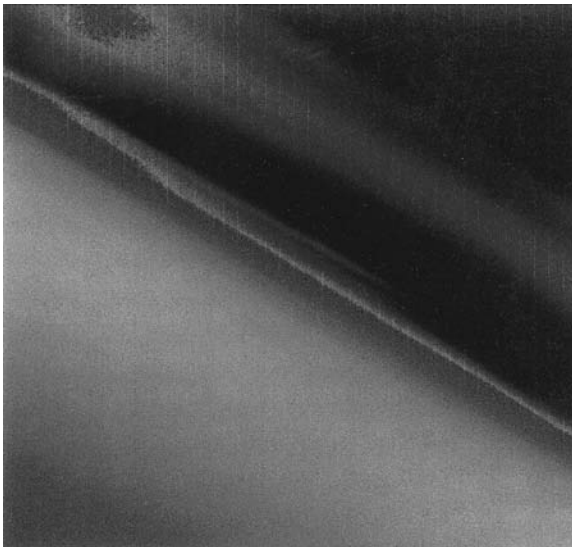


Figure 9. Voyager 1 picture of Titan's limb showing the extensive high altitude haze extending to 700 km. The colors are false and are chosen to bring out the details of the haze structure (NASA picture). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

be hoped. The basic parameters of Titan's physical properties and its atmospheric structure and composition are given in Tables 6 and 7.

The Voyager experiments found that CH_4 is indeed a lesser constituent of Titan's atmosphere, estimated at roughly 4% with a possible range of 2–10%. Thus, the exact amount of CH_4 column abundance, although much better confined, is still not precisely determined. The major constituent of Titan's atmosphere is molecular nitrogen (N_2), just as on Earth. The surface pressure on Titan is 1.5 atm., and the surface temperature is 94 K. Although this surface pressure is only slightly greater than that of Earth, the amount of Titan's atmosphere in numbers of molecules is actually 11 times greater. Because of Titan's lower surface gravity (about seven times less than that of Earth), a greater amount of atmosphere is needed to provide the same surface pressure. The lower surface pressure also makes Titan's atmosphere very much more extended than that of Earth. Its tropopause is at 42 km, whereas for Earth, it is around 10–12 km. Titan's atmosphere from the surface to the visible cloud tops, near 1 mb pressure extends a full 200 km. This makes the visible pressure extends a full 200 km. This makes the visible disk seen in a telescope considerably larger than the solid surface.

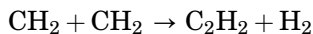
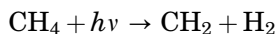
The reason for Titan's diverse composition can be found in the extraordinary photochemistry of its atmosphere. At high altitudes, methane is dissociated by the energy of solar ultraviolet radiation and forms free radicals such as CH_2 , CH_3 , and atomic hydrogen. The very reactive radicals form more complex hydrocarbon molecules such as ethane (C_2H_6), propane (C_3H_8), ethylene (C_2H_4), and acetylene (C_2H_2). The set of reactions that forms these molecules is rather complex. It includes creating and destroying radicals, vertical and horizontal

Table 7. Gases Observed in Titan's Atmosphere

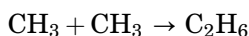
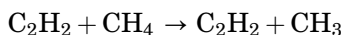
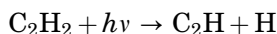
Species		Abundance ^a
Nitrogen	N_2	~95%
Methane	CH_4	~4%
Hydrogen	H_2	0.1%
Carbon monoxide	CO	~50 ppm
Ethane	C_2H_6	15 ppm
Propane	C_3H_8	0.9 ppm
Acetylene	C_2H_2	5 ppm
Ethylene	C_2H_4	~7 ppm
Hydrogen cyanide	HCN	0.8 ppm
Methyl acetylene	C_3H_4	~21 ppb
Diacetylene	C_4H_2	~14 ppb
Cyanoacetylene	HC_3N	~0.05 ppm
Cyanogen	C_2N_2	~0.02 ppm
Carbon dioxide	CO_2	0.014 ppm
Acetonitrile	CH_3CN	Detected
Dicyanoacetylene	C_4N_2	Solid phase

^appm: parts per million mean polar/equatorial mole fraction; ppb: parts per billion; ~: approximate abundance.

transport in Titan's atmosphere, and collisions with the abundant constituents of Titan's atmosphere. A simplified set of reactions that produces acetylene and ethane using solar or photon energy, ν being the frequency of radiation, and h Planck's constant, is illustrated below.



The acetylene produced can now form reactions that make ethane:



Because of the relatively low surface gravity of Titan, atomic and molecular hydrogen escape. Thus these reactions are not reversible, and Titan loses hydrogen steadily at a rate of roughly 5×10^{27} atoms/s. This loss of hydrogen and its replenishment by photochemical means results in an equilibrium concentration of 0.2% H_2 in Titan's atmosphere. Note that the photochemistry described is unique to Titan. It does not occur on Saturn because that planet has a huge reservoir of H_2 which is held tightly by its very much larger gravity field.

The molecules produced include nitriles which are organic molecules that contain nitrogen. Additionally, oxygen-bearing molecules have been observed. The nitriles are formed by dissociation of N_2 which has a very strong bond and is hard to break, so that it requires very energetic short wavelength UV radiation or electron impact dissociation high in Titan's atmosphere. Chief among the nitriles are hydrogen cyanide, HCN and cyanogen (C_2N_2). The oxygen-bearing molecules are believed to come from impacts by comets, that contain $\sim 50\%$ water and can thus supply oxygen.

The photochemistry of Titan has been verified both by theoretical calculations and laboratory simulations. Essentially all of the molecules observed in Titan's atmosphere have been produced in roughly the proper abundance, using laboratory simulation experiments. The net result of Titan's photochemistry is the loss of hydrogen and the production of heavier and more complex hydrogen-deficient organic molecules. As of the year 2000, 13 complex molecules were identified in Titan's atmosphere, as listed in Table 7. Laboratory simulations have produced an additional 20 or so complex hydrocarbon molecules that should be present in sufficient abundance that they should be detectable by some of the instruments of the *Cassini spacecraft Huygens probe*.

There are further important consequences of Titan's photochemistry, which explain the satellite's pervasive haze layer. Complex organic molecules can connect to form long-chain polymers. These polymer chains can apparently grow to complexes containing several hundred molecules and reach a size of the order of

0.1 μm radius. These small particles in turn form irregular aggregates of 0.4–0.5 μm radius, and these aggregates form aerosol particles in Titan's atmosphere. Their approximate size has been determined by their light scattering properties from the ultraviolet to the infrared.

The present consensus holds that the haze in Titan's atmosphere does not extend to the surface, but that the atmosphere below the tropopause at a height of about 70 km is clear. Calculations further show that it would take 50 years for the 0.5- μm smog particles to fall to the surface, so that some mechanism must exist to clear the atmosphere of this haze, very possibly condensation and precipitation of methane rain.

Whatever the mechanism, a large amount of organic material, be it aerosol particles or condensible gaseous species such as ethane, must have accumulated for possibly billions of years, and it still accumulates on Titan's surface. What the surface of Titan looks like, is therefore, an extremely tantalizing and an intriguing question. Glimpses of Titan's surface have been obtained by imaging in regions of minimum CH_4 absorption (so-called windows in the CH_4 spectra) using the Hubble space telescope. These images reveal continent-sized albedo variations, but more detailed conclusions cannot be drawn. There is speculation about liquid methane lakes or oceans (a continuous source of methane in the atmosphere is required to replenish the methane that is being destroyed by photochemistry) or oceans several hundred meters to kilometers deep that contain a mixture of ethane, methane, and nitrogen, or possibly there is only dry land, and mechanisms that reprocess the organic material on the surface have not presently been considered.

Titan is an object in our solar system, which despite spacecraft visits, still has a large number of extremely interesting unanswered questions. For this reason, the Cassini mission decided to send a special probe called Huygens into Titan's atmosphere, rather than the atmosphere of Saturn. It is hoped that this probe will answer the many questions about Titan's unusual character.

One of the big mysteries of our Earth is, of course, how life began. A number of models of early Earth's atmosphere postulate an extensive primordial atmosphere made up of H_2 , CH_4 , NH_3 , N_2 etc. similar to the major planets. Through interactions with an intense solar wind, the Earth's hydrogen was soon lost and left behind an atmosphere similar to that on Titan. The process of photochemistry could then produce the same suite of complex organic molecules that we see on Titan. Among these molecules, nitriles are the most important, because such molecules as hydrogen cyanide (HCN), cyanogen (C_2N_2), and cyanoacetylene (HC_3N) play a crucial role in forming amino acids which are believed, to be the precursors to simple cell formation. Once life begins, photosynthesis produces oxygen. In a sense, Earth has an atmosphere very similar to that of Titan, in which the main constituent is N_2 . The major differences lie in the second major constituent, which on Titan is the reducing gas CH_4 , whereas on Earth it is the oxidizing molecule O_2 . Some of the O_2 on Earth is produced by photodissociation of H_2O , but the majority comes from plant photosynthesis. This process started after life began on Earth. On Earth, the surface temperature is, of course, much higher than on Titan, allowing reactions that would not occur on this satellite. Perhaps if the temperature were 50–100 degrees higher on Titan, life could form on this satellite also. In any case, the satellite allows us to study the type of

chemical reactions and photochemistry that is analogous to the prebiotic processes that were at work on primitive Earth.

UWE FINK
Lunar and Planetary Lab
University of Arizona
Tucson, Arizona

SCIENCE FROM SOUNDING ROCKETS

Introduction

While researching this article, I had the opportunity to visit the White Sands Missile Range to witness the launch of a scientific payload on a Black Brant rocket. The payload contained a new kind of optical telescope that could yield information about the atmosphere of a white dwarf star. The rocket took off without incident at its scheduled time (about 10 PM local time). However, within a minute, it was apparent that something was wrong. The rocket was drifting to the west and was projected to land off the range, violating safety rules; thus the flight controller was forced to initiate a small explosion in the rocket and cut the flight short. High-altitude winds had suddenly shifted in the 30 minutes before the flight, and the wind corrections, which had been entered just moments earlier, were not accurate. Fortunately, the instrument was recovered the next morning in excellent condition, so that it could fly again at a later date. So, even after 50 years, scientific rocketry remains a high-risk enterprise.

Everyone at White Sands that day witnessed the conjunction of two grand traditions—the striving of scientists to conduct research at increasingly higher altitudes and the development of the technology to carry instruments high into the atmosphere. In this article, the focus is principally on rockets that carry instruments for short durations, up to 5 minutes or so, before falling back to Earth. These rocket flights are also described as suborbital because the rockets do not achieve the high velocity needed to remain in an orbit around Earth.

Rockets that carry experiments are typically called “sounding rockets.” “Sounding” now means any observation of the properties of the ocean or atmosphere. Originally the term applied to the technique used to measure the depth of the water under a ship with a sounding line. The use of “sound” probably derives from old English, where “sund” meant sea. One of the first scientific uses of rockets was to measure the properties of the atmosphere, so naturally they were given the name sounding rockets, which has stuck.

High-altitude research has a long history. Blaise Pascal conducted a high-altitude experiment in the 1640s. To demonstrate that air had weight, he had a Toricelli barometer carried up the Puy de Dome, an extinct volcano 1400 meters high. The importance of carrying scientific instruments to high altitudes for

observing through the atmosphere was expressed eloquently by Sir Isaac Newton (1). In discussing the development of telescopes for astronomical purposes, he notes the limitation presented by the “twinkling” of the stars and writes, “The only remedy is a most serene and quiet air, such as may be found at the tops of the highest mountains, above the grosser clouds.” By 1900, scientists were more or less routinely making expeditions to mountaintops and carrying instruments in balloons to conduct observations.

Rocketry has an even more ancient history, beginning with the discovery by the Chinese of gunpowder and its use in bamboo tubes in the thirteenth century to create projectiles of a sort. From a military perspective, gunpowder found its great utility in muskets and cannons, and it was not until around 1800 that practical rockets were used in warfare (and entered U.S. consciousness with the British bombardment of Fort McHenry during the War of 1812). Rocketry, or at least the human exploration of space, received a tremendous boost from the writers of the nineteenth century, especially Jules Verne, whose book, *From the Earth to the Moon*, was widely read.

Science and rockets came together in the twentieth century. Major discoveries were made at the beginning of the century relating to high-altitude phenomena; specifically, the stratosphere, which demonstrated that the atmosphere was structured; cosmic rays, which were energetic radiation originating high in the atmosphere; and the long range transmission of radio waves that revealed the existence of a conducting layer in the atmosphere at high altitude. These and other advances led to pressure from the science community to perfect the means for increasingly sophisticated experiments at high altitude. Meanwhile, three extraordinarily talented physicists—Konstantin Tsiolkovsky in Russia, Hermann Oberth in German, and Robert Goddard in the United States—established the theoretical and technical basis for rockets. Unfortunately, it took a war and the search for an ultimate weapon by the Nazis to bring this all together in the development of the V-2 rocket by Dornberger and his team at Peenemunde. After World War II, scientists in the United States capitalized on the captured inventory of German rockets to conduct a variety of scientific observations and to develop rockets of their own, which formed the basis for the space science program in the United States and elsewhere.

Three sources were used principally in preparing this article: sites on the Internet, books, and original articles. The books include the following:

- Frank H. Winter, *Rockets into Space*, Harvard University Press, 1990 provides a summary of the history of the development of rockets in the twentieth century and its prehistory.
- David H. DeVorkin, *Science With A Vengeance*, Springer-Verlag, 1992, provides a detailed account of the emergence of rocket science following the development of the V-2 rocket.
- Bruce William Hevly, *Basic Research Within a Military Context: The Naval Research Laboratory and the Foundations of Extreme Ultraviolet and X-ray Astronomy, 1923–1960*, University Microfilms International, 1992, describes high altitude and rocket research at one of the centers where rocket science emerged.

- Homer E. Newell, *High Altitude Rocket Research*, Academic Press, 1953, provides a summary of rocket technology and of the status of the various science disciplines that used rockets up to about 1952.
- R.L.F. Boyd and M.J. Seaton (eds), *Rocket Exploration of the Upper Atmosphere*, Pergamon Press, 1954, describes the state of the science emerging from the use of sounding rockets up to mid-1953.

Only the Winter and DeVorkin books are likely to be found in a general purpose library. The Newell and Boyd/Seaton books are likely to be found only in major research libraries. The Hevly book is his Ph.D. thesis and is of even more limited availability.

The discussion cited later of foreign rocket programs, information on the characteristics of individual rockets, a description of Pascal's experiment with a barometer and of the Montgolfiers early balloon flights, a biographical memoir from Frank Malina—all these and much more information came from the Internet. I do not cite specific sites because they tend to be short-lived, but others take their place. My recommendation to the reader is to use one of the many web search engines with a few key words (e.g., Pascal and barometer) or an exact expression (e.g., discovery of ozone in the atmosphere).

The Development of Rockets

The physics and technology of rockets are discussed in other places in this Encyclopedia. However, for the nontechnical reader, I note the following. The motive force for rockets is produced by expelling material, usually hot gas, at high velocity. This force is the result of Newton's third law stating that for every action, there is an equal and opposite reaction. More precisely, it is Newton's second law that shows that a forward force is produced on the rocket equal to the product of the rate at which mass is expelled and its velocity. This is in contrast to a gun in which the gas is contained in a barrel. The expansion of the hot gas in the gun's barrel provides the forward impulse to the bullet. In fact, a key experiment performed by Robert Goddard, the American pioneer of rocketry, was to demonstrate that the performance of a rocket in a vacuum was better than that of a rocket in air. The extra "push" caused by the expansion of the exhaust gas against the external medium is more than compensated for by the greater exhaust velocity of the expelled gas if it encounters no external resistance. So in rocket weapons, such as the American Bazooka or the Russian Katyusha, the barrel is left open in the backward direction to allow the exhaust gases to escape freely. Beyond this simple principle lies an enormous amount of technology, starting with the rocket engines needed to produce the high-velocity gas to the guidance systems that ensure controlled flight.

At the beginning of the twentieth century, rocket technology did not exist, nor was Newton's third law generally recognized as the guiding principle of rocket propulsion. Even as late as 1920, the New York Times editorialized,

As a method of sending a missile to the higher, and even to the highest parts of the earth's atmospheric envelope, Professor Goddard's rocket is a practicable and

therefore promising device. It is when one considers the multiple-charge rocket as a traveler to the moon that one begins to doubt ... for after the rocket quits our air and really starts on its journey, its flight would be neither accelerated nor maintained by the explosion of the charges it then might have left. Professor Goddard, with his "chair" in Clark College and countenancing of the Smithsonian Institution, does not know the relation of action to re-action, and of the need to have something better than a vacuum against which to react ... Of course he only seems to lack the knowledge ladled out daily in high schools. (2)

As frequently happens, the science and technology of modern rocketry arose independently in several places. Frank R. Winter's book, *Rockets into Space* (3), provides a summary of this early history. Konstantin Tsiolkovsky is generally given credit for the earliest work that led to contemporary rockets. He was born in Izhvesk, far to the east of Moscow, and was educated in Moscow. He spent his professional career in Kaluga, a town 160 kilometers from Moscow. He developed with mathematical rigor many of the basic relationships of rocket motion, for example, a rocket's velocity as related to the exhaust velocity of the propellant. Although highly regarded as a scientist within the Soviet Union, he was elected to their Academy of Sciences in 1919, he seems to have had little influence on rocket development, even within his own country. He was essentially unknown in the West. It was not until 1940 that his works appeared in an English translation. Also, he did not actually build rockets or conduct experiments, nor was he associated with any groups that did.

Hermann Oberth, the pioneer of German rocketry, was born in Rumania to ethnic German parents. By the age of 15, he had designed his first rocket and in his early 20s, in 1917, had proposed to the German War Department the development of a rocket strikingly similar to the V-2. (The proposal was rejected.) In 1923, he published "Die Rakete zu den Planetenraumen" (Rockets to Outer Space), an outgrowth of his Ph.D. thesis. The book was very broad in its treatment of rockets and rocket flight and included problems related to manned space flight and to conducting science from spacecraft. In addition to attracting wide attention, it may have been the trigger for the international astronautics movement. Winter (3) remarks that because of his wide influence, Oberth deserves the title "Father of the Space Age."

However good Oberth's ideas were, his rockets existed only on paper. Robert Goddard converted his own ideas to practice. He was born in Worcester, Massachusetts, about 100 kilometers west of Boston, educated in Worcester, and spent his professional career there. Goddard developed rockets through his interest in spaceflight. His diaries, beginning when he was 17, reveal a tortuous process, exploring every conceivable means of rocket propulsion until he arrived at the liquid-oxygen-fueled rocket in 1909 while a graduate student at Clark University. In that same year, he performed his first experiment related to rocket propulsion. By 1916, his work had advanced to the point where he was able to seek and receive support for rocket development from the Smithsonian Institution in Washington. His goal at that time was to produce a rocket that could carry instruments significantly higher than balloons could. Goddard achieved notable technical success with his research, but the performance of his rockets was modest compared to the much better supported German effort. The U.S. military

never put significant resources into long-range rockets either before or during World War II. Goddard's own wartime contribution was to the development of the JATO rockets, used to assist aircraft at takeoff.

Science at High Altitude

The nineteenth century witnessed the Industrial Revolution in which mechanical and electrical engines replaced humans, water, and wind as the principal motive forces. A similar revolution took place in science and exploration. Using the tools provided by the new industries, the scientific methodology developed by previous generations and their national wealth, investigators developed new understanding of the natural world and made astounding discoveries of phenomena never before observed. The new technology allowed the development of increasingly sophisticated instruments; it also allowed scientists to conduct expeditions far removed from their laboratories and even to leave the surface of Earth if necessary. One of the principal tools for high altitude research after 1900 was the balloon. It had its start in France more than a century earlier when two papermakers, Joseph and Etienne Montgolfier, noticed that burning paper and smoke rose up the chimney and wondered if they could apply that phenomenon to build a flying machine. They built their first balloon of paper and fabric and fueled it with a very smoky fire without understanding that it was the hot air, not the smoke, that caused the lift. The first manned flight occurred in Paris on 21 November 1783 before a crowd of 400,000. Soon afterward, the first sealed gas balloon was flown, using hydrogen. Ballooning became a way to perform high-altitude research and was a stepping-stone to science in outer space. In 1901, Reinhard Suring and Arthur Berson of the Royal Meteorological Institute in Berlin flew to an altitude of 10.5 km, still the record for an unmanned, open balloon flight. The flights of the Piccard brothers in sealed capsules during the 1930s were especially well publicized. Auguste Piccard, a particularly eloquent scientist, described the scene thus:

At an altitude of ten miles, the Earth is a marvelous sight. Yet it is terrifying, too. As we rose the Earth seemed at times like a huge disk, with an upturned edge, rather than the globe that it really is. The bluish mist of the atmosphere grew red-tinged, and the Earth seemed to go into a copper-colored cloud. Then it all but disappeared in a haze (4).

Effect of the Atmosphere on Scientific Investigations

Astronomers were among the first scientists to recognize the disturbing influence of the atmosphere on observations. Although the sky can appear perfectly transparent, it creates a number of deleterious effects. For starters, it must be remembered that there is a lot of material above us, amounting to about ten thousand kg/m². It is remarkably transparent to visible light from blue to red, considering that the equivalent weight in glass would be about 4 meters thick. However, absorption becomes significant outside this band of colors in both the ultraviolet and infrared radiative bands. It is not the dominant constituents, nitrogen and oxygen, that are responsible for the absorption. In the infrared, the

culprits are water vapor and carbon dioxide, better known now as greenhouse gases. In the ultraviolet, the culprit is ozone; it is the decrease of ozone as an atmospheric constituent and the subsequent increase of ultraviolet radiation at Earth's surface that is of current concern to the climate change community. Further into ultraviolet (shorter wavelength), the atmosphere becomes highly opaque to radiation, as it is for particles that comprise cosmic rays.

Astronomers face absorption in the atmosphere and also refraction that causes light to bend. The degree of bending is a minor problem; however, small variations in density in the atmosphere make the bending very unsteady. To the human eye or to a small optical telescope, these changes cause images such as stars to move around and vary in brightness, to "twinkle." In a large telescope, the images enlarge beyond what the performance of the optics might allow.

Scattering is another effect of the atmosphere on radiation. Our blue sky is the result of scattering of sunlight, as is our red sunset. The effect of scattering is particularly noteworthy following a major volcanic eruption. The volcanic ash spewed into the atmosphere can result in brilliantly colored sunsets and sunrises for months or even years after such an event. Scattering is especially serious when trying to observe a faint object in the presence of a bright object. The scattered light from the bright object can overwhelm the brightness of the faint one.

Atmospheric Science

Labitzke and van Loon provide a thorough discussion of the history of investigations of the atmosphere (5). By 1900, the idea had been dispelled that the atmosphere was simply a gaseous mix of nitrogen and oxygen whose density and temperature decreased as one went to higher altitude. In 1896, Teisserenc de Bort, flying unmanned, instrumented balloons from Trappes near Versailles, discovered that after falling continuously with increasing altitude, the temperature began to rise above 11 km. De Bort, recognizing this sudden rise in temperature as evidence that the atmosphere was layered, coined the terms troposphere and stratosphere to describe the two layers. The discovery of ozone in the atmosphere was another great success of the early twentieth century. It had been found that solar radiation cutoff sharply in the ultraviolet, below about 3000 Å. In 1880, W.N. Hartley proposed that the cutoff results from absorption by ozone in the atmosphere. The actual distribution of ozone high in the atmosphere was demonstrated by Lord Raleigh in 1917, based on observing the rising and setting sun. Then, in a remarkable precursor to experiments from rockets, Erich and Victor Regener measured the solar spectrum in a balloon, as it rose to a peak altitude of 31 km and fell back through the atmosphere. They were able to trace out the distribution of the ozone directly through its effect on the Sun's radiation.

Ionospheric Investigations

Guglielmo Marconi was responsible for one of the astounding developments of the period when, in 1901, he transmitted a radio signal across the Atlantic Ocean. Within a few years, "wireless telegraphy" was being used for practical

communications across long distances on Earth. Also within a few years, it was recognized that there had to be a layer at high-altitude above Earth that could reflect radio waves and permit them to travel long distances. Otherwise, the radio waves would escape into space, and reception would be limited to 50 miles or so, depending on the height of the receiving and transmitting antennae. Oliver Heavyside wrote in 1902,

There may possibly be a sufficiently conducting layer in the upper air. The guidance (of radio waves) will then be by the sea on one side and the upper layer on the other. (6)

It was more than 20 years before scientists traced out the characteristics of this “conducting layer,” specifically, that it was composed of electrons at high altitude. Robert A. Watson-Watt coined the term ionosphere to provide a continuous sequence with troposphere and stratosphere, the other layers in the atmosphere.

Radio propagation and the ionosphere attracted the attention of military organizations, none more so than the U.S. Navy because of its responsibility for maintaining a global fleet. Fortuitously, the Navy had created the Naval Research Laboratory in Washington, D.C., during the 1920s. Radio propagation was one of its major concerns, where it attracted the attention of E.O. Hulburt. Hulburt had broad interests as a scientist; his first research on the subject dealt with the characteristics of the ionosphere as a medium from which to derive radio propagation effects and vice versa. But then he asked the question, what caused the electrons to be there, and decided that ultraviolet radiation from the Sun was the likely culprit. It was more than just an inspired guess because it was believed that the Sun’s corona, well studied from eclipse observations, had to be composed of hot gas that would radiate copious amounts of ultraviolet radiation. Meanwhile, across town from the Naval Research Laboratory, at the Carnegie Institute, Gregory Breit and Merle Tuve had developed the technique of studying the ionosphere by radio sounding, sending up pulses of radio waves and watching for their return. This layer of the atmosphere was well above the altitude accessible to balloons. Hulburt, Breit, and Tuve knew of Goddard’s work on rockets and had discussed the possibility of using rockets to conduct observations at high altitude, but nothing much came of their interest until after World War II.

Cosmic Rays

Bruno Rossi discussed the early history of cosmic rays and the role of high-altitude research in developing the discipline (7). Natural radioactivity was discovered in the laboratory by Henri Becquerel in 1896. By 1900, its distribution was being widely studied, including from above the surface of Earth. Scientists reasoned that if radioactive material was distributed throughout Earth’s crust like other minerals, the resulting radiation should decrease away from the surface because of attenuation in air. Such was the case as one left the surface, but then, inexplicably, the intensity of radiation began to increase. In 1912, Victor Hess, carrying his instruments in balloons floating above Austria, found that the

intensity was four times greater at an altitude of 16,000 feet than at the surface. Hess hypothesized that there had to be a source of energetic particles falling upon Earth from above the atmosphere. A decade later, Robert A. Milliken proved the hypothesis to the satisfaction of the scientific community, based on measurements in mountain lakes in California. Milliken demonstrated that the intensity of the radiation at the bottom of such lakes was reduced by an amount corresponding to the absorption in the overlying water. It was Milliken who coined the term “cosmic rays” for this new phenomenon. However, the radiation still had to penetrate more than half the atmosphere to reach even the highest mountains. Thus, there was little likelihood that the primary radiation impinging on the top of the atmosphere was actually being observed. So, to learn its true nature, it was necessary to get as high as possible in the atmosphere, and scientists continued to put instruments into high-altitude balloons. In one of the last prewar cosmic ray experiment in the United States, measurements were obtained in 1940 by scientists at the University of Chicago at an altitude of 70,000 feet, where only 3% of the atmosphere remains.

Astronomy and Solar Physics

Until fairly late in the nineteenth century, it was still common for astronomers to build observatories in or near their home institutions, even if it was a major city. However, in 1874, when James Lick’s Board of Trust in San Francisco began to consider the best possible site for an observatory, they investigated mountaintop sites despite the logistical problems and chose Mount Hamilton whose elevation is 4200 feet. Lick’s observatory (appropriately the Lick Observatory) became the first permanently occupied mountaintop observatory in the world. Although city lights were an important issue, the improved clarity of the sky was impressive. In 1909, a scientist at the French observatory on 2.9-km high Pic du Midi in the Pyrenees commented,

“The sky suddenly clears and I have before my eyes the most unimaginable scene that an astronomer can dream of. The Milky Way is sparkling, the stars shine like beacons. The sky is white with stars and their brightness is enough to light up the clouds that are at our feet.”

It also became apparent to astronomers that the Sun and the stars were producing radiation in the ultraviolet and at shorter wavelengths that was unobserved at Earth’s surface because of the opacity of the atmosphere. By 1900, the general nature of stars was well understood; in particular, it was known that the stars radiated because their surfaces were hot. In the case of the Sun, the temperature is about 5700 K and its radiation spreads from the red to the blue, peaking at about 4500 Å in the green color range. However, observations of stars indicated that they existed in wide range of surface temperatures, exceeding 10,000 K. Most of the radiation of such hot stars would be in the ultraviolet range and would be absorbed in the atmosphere.

During the 1930s, the remarkable discovery was made that the temperature of the gas in the Sun’s atmosphere was several millions degrees. This was based on identifying spectral lines in the visual range from very hot atoms of

elements such as iron and calcium. The lines themselves had been discovered more than 50 years earlier during eclipse observations. Such hot gas implied the emission of X rays from the Sun.

The Development of the V-2 Rocket

Ever since siege guns spelled the end of fortified cities, long-range artillery has been a staple of military inventories, certainly in Europe. Thus, the German Army recognized the potential of rockets as the basis for a new kind of cannon and in 1932 began their developing at a test station in Kummersdorf near Berlin, under the direction of a young Army officer, Walter Dornberger, who has written a memoir covering the development of the V-2 (8). Among his assistants was Wernher von Braun, then a 19-year-old student at the Berlin Institute of Technology. Their first project, a 650-pound thrust rocket engine, blew up. (Thrust, the standard measurement of rocket engine performance, is the product of the rate at which mass is expelled from the rocket and its velocity. By Newton's laws, it is also the forward force on the rocket. It is expressed in pounds [English units] or Newtons [metric units]. One pound of thrust equals about 4.5 Newtons.) However by 1936, they had developed a working engine that had a thrust of 3500 pounds (the highest Goddard achieved was 825 pounds in 1941; the V-2 would attain a thrust of 55,000 pounds) and had attracted the attention of the German General Staff. With ensured support, the operation was transferred to Peenemunde on the Baltic coast where it would remain through World War II. Although well supported, the program did not receive the highest priority for weapon development until 1943. The first V-2 rockets were fired at London in September 1944.

Dornberger's group recognized that they did not have adequate information about the atmosphere at high altitude to calculate the trajectory of their rockets. In 1942, Von Braun and his staff briefed academic scientists around Germany on the program, including Erich Regener. Regener had headed a physics institute at Stuttgart and was a leading German scientist; his personal research involved using balloon-borne instruments to study cosmic rays and the properties of the upper atmosphere. His discovery of the ozone layer has already been noted. He did not have an easy time in Nazi Germany because he opposed the idea of Aryan science and had a Jewish wife. His son had fled Germany in 1936 and he himself lost his position at Stuttgart. By 1942, he had reestablished himself in Berlin and was continuing his research, supported in part by the German Air Ministry. He then began the development of instruments for the Peenemunde group that would fly on their rockets to measure the temperature, pressure, and absorption of solar radiation by ozone. In addition, he included an instrument for the German astronomer, Karl Kippenheuer, to measure solar ultraviolet radiation directly. Regener's idea was to have these instruments contained in a capsule to be ejected from the rocket at its peak altitude and float to earth on a parachute, taking measurements on the way down. In spite of wartime pressure, the capsule and its instruments were fabricated and ready for installation in V-2 rockets by the end of 1944. But before preparations could be completed, the advancing Russian army forced the evacuation of Peenemunde.

It has been estimated that the development of the V-2 was as costly to the Germans as the atomic bomb was to the Americans. In addition to the facilities at Peenemunde, tens of thousands of concentration camp inmates, forced laborers from foreign countries, and German workers were employed in manufacturing the high-priority missiles at the Mittelwerk GmbH camp near the town of Nordhausen (Thuringia). It was certainly one of the great technical achievements of modern times.

Transition to the Postwar Era

The period between 1900 and 1945 set the stage for the scientific revolution that took place after the end of World War II, based on the use of rockets for scientific purposes. Scientists had developed compelling objectives in a variety of disciplines that required observations from high altitude. They had also developed the tools to do so. Of greater significance was that the scientists of the prewar period were still active in 1945. They, along with the scientists who emerged from the wartime laboratories, were instrumental in creating the revolution. DeVorkin provides an detailed discussion of the postwar period of using sounding rockets for research purposes up to about 1955 (9).

Even before the end of the war, U.S. agencies were in Europe searching out information about German science and technology. The dramatic effect of the V-2 bombardments made rocket technology one of the focal points of interest. The U.S. Army Ordnance Department retrieved tons of V-2 parts and documentation from the Nazi factories at Mittelwerk and shipped them back to the United States. Ernst Krause from the Naval Research Laboratory and member of a Navy group in Europe returned with the knowledge that sophisticated rockets could be built. The Alsos project, an effort focused on Germany's atom bomb development, included Gerhardt Kuiper, an astronomer who had extensive experience in radio science. Kuiper learned about Regener's wartime efforts to fly instruments on the V-2 and other upper atmospheric research in Germany. He transmitted that information to Donald Menzel at Harvard who distributed the information to other U.S. scientists, notably to Merle Tuve, E.O. Hulburt, and Leo Goldberg, then at the University of Michigan. Another activity, CIOS (Combined Intelligence Objectives Subcommittee) included Fritz Zwicky from the California Institute of Technology who came back a space enthusiast. His own desire was to study the effects of hypervelocity at high altitude by creating artificial meteors.

Following the end of the war in Europe, the United States engaged in a furious effort to assemble and fly V-2 rockets. The U.S. Army Bureau of Ordnance established a capability to rebuild V-2 rockets, having installed Von Braun and other members of the Peenemunde team at Fort Bliss in Texas. The rockets would then be flown from the newly established missile range at White Sands, New Mexico, under the direction of Colonel Holger Toftoy. Their primary directive was to gain experience in handling and firing missiles; their next objective was to conduct high-altitude research. To accomplish this latter objective, Toftoy sought out the nation's military and civilian scientific organizations. Toftoy was no doubt impressed by the wartime partnership between the scientific and

military communities and so naturally turned to scientists again to accept a role in this new endeavor. It is important to remember that in 1945, the United States did not have a formal process to conduct scientific research; both the National Science Foundation and NASA were more than a decade away. So individual agencies made individual arrangements. The Naval Research Laboratory convinced the Navy's Office of Research and Invention, the forerunner of the Office of Naval Research, that it should have the lead in the Navy for conducting rocket research. The Applied Physics Laboratory, established during World War II under Merle Tuve, received sponsorship for rocket research and rocket development from the Navy's Bureau of Ordnance. They were the major players, but other military agencies and major universities joined in. After much haggling, well described by DeVorkin (9), the V-2 Rocket Panel was formed with membership from NRL, APL, Cal Tech, Harvard, University of Michigan, and other organizations to oversee the allocation of space on V-2 rockets for high-altitude research. The research goals included radio and sound propagation in the atmosphere, properties of the atmosphere, cosmic rays, solar ultraviolet radiation, and various biological investigations. The first flight of a V-2 took place on 16 April 1946. The last V-2 was flown in 1952. By then, the Aerobee rocket, first flown in November 1947, had become the principal rocket used by American scientists. Figure 1 shows an assembled group of military personnel at White Sands preparing a V-2 for flight in 1947.

The Aerobee rocket was the culmination of work on rockets that had begun during the 1930s at the California Institute of Technology within a group headed by Frank Malina that eventually would become the Jet Propulsion Laboratory. Malina, then an engineering graduate student, was part of a group headed by Theodore Von Karmen, the great aerodynamicist. By 1939, the group obtained support from the Army Air Corps for JATO development. In January 1945, Malina and his colleagues proposed to develop a rocket that would carry a 25-pound payload to an altitude of 20 miles. In October 1945, they launched the first flight of their rocket—the WAC Corporal. It reached an altitude of 45 miles. It was America's first sounding rocket, even though it could not compare in performance with the V-2. As an aside, Malina, in a biographical memoir, traces his interest in space exploration to reading Jules Verne's, *From the Earth to the Moon*, when he was a boy of 12.

The WAC Corporal was built by the Aerojet Corporation, a private company that had emerged from the Cal Tech group in 1941 and built thousands of JATO units during the war. They were then asked by the Applied Physics Laboratory to build a new rocket that would carry 150 pounds to 50 miles and would become the Aerobee. The first one was launched in November 1947 and remained in service, with various modifications, through the 1980s. The Naval Research Laboratory, also recognizing the need for high-performance rockets, contracted with the Glenn L. Martin Company to build the Viking rocket to reach an altitude of 150 miles. The first Viking was flown in May 1949. It was too expensive to achieve wide use as a sounding rocket, but it did form the basis for the Vanguard program, the U.S. national program to put a satellite in Earth orbit. A remarkable aspect of these developments was their rapidity. In the case of the WAC Corporal and the Aerobee, less than a year was required to develop and fly the first rockets.

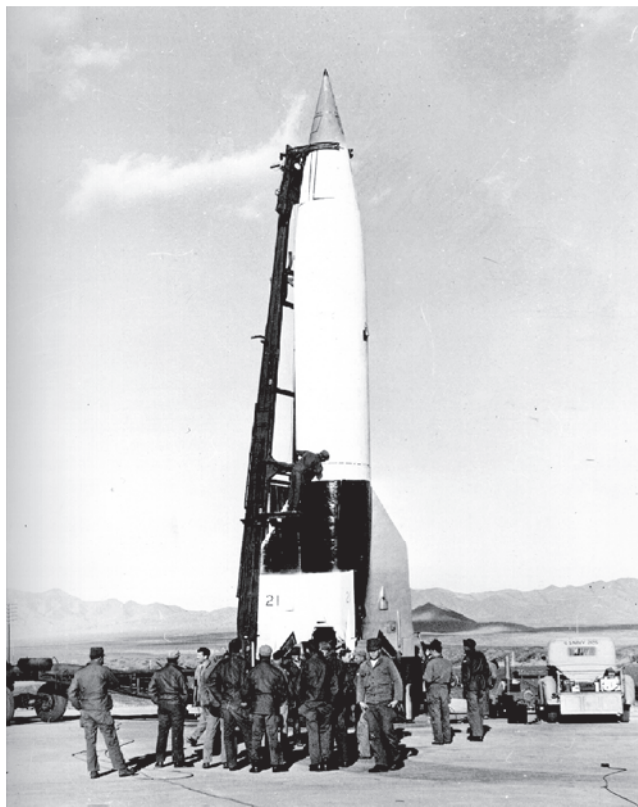


Figure 1. A V-2 rocket being prepared for flight at the White Sands Missile Range in 1947. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The importance of rockets for military and science applications was apparent to countries other than the United States. Work on rockets began in the United Kingdom in 1946 within the Royal Aircraft Establishment at Farnsworth. In 1955, U.K. scientists promoted the development of a rocket, which was to become the Skylark, that could lift 45 kg to 210 km. The first flight of the rocket took place in Woomera, Australia, during February 1957. The Skylark and the Woomera range have become staples in the British high-altitude research program. The Skylark has also been used by the Germans and the Swedes in their national research programs. France also began rocket development in 1946 in its Ballistics and Aeronautical Research Laboratory (LRBA) with first the EOLE, then Veronique, Vesta, Belier, Centaure, and Dragon. Launches from Hannaguir, Algeria, began in 1952.

Now, the U.S. sounding rocket program is managed by NASA and uses of thirteen different configurations, most of which are multistage. The earliest rockets fall into the lowest performance category of the current capability. In spite of the vast improvement in performance, science from rockets is still limited in time available for experimentation, which is intrinsic in the ballistic nature of the flight.

Rocket Performance

The first rocket scientists were primarily interested in achieving some level of thrust and altitude with their rockets. With the advent of the V-2, scientists and engineers had to achieve specific operational requirements, namely, to carry a 2300-pound warhead a distance of about 190 miles. To achieve that goal, the rocket would have to reach an altitude of about 60 miles. If there were no atmosphere, the rocket could achieve the 190-mile distance and reach an altitude of only about 48 miles. Atmospheric resistance requires that the rocket be fired more directly up than optimum, so that it spends less time in the lower atmosphere.

To conduct science, the V-2 and other rockets could be launched straight up. Because rocket ranges tended to be small, it was required that the rocket not fly very far from its launch site. For scientific purposes, V-2s achieved peak altitudes of more than 100 miles. For vertically launched rockets, the time in orbit depends only on the peak altitude, and the observation time for a given experiment was given by the altitude range in which data could be obtained. The size of rocket ranges and the cost of the rockets have limited their performance, so that even the latest rockets do not exceed about 300 miles in altitude and provide observing time of about 5 minutes for payloads of about 1000 pounds.

There has been one exception to this. After World War II, the range requirement of military missiles went from hundreds of miles to thousands of miles. To meet this requirement, the United States fired rockets from a range in California to islands in the South Pacific. These rockets allowed about 30 minutes of observation time. During the 1960s, a group at the Lawrence Livermore National Laboratory took advantage of these rockets to conduct astronomical observations (10).

Early Science Using Rockets

Science using sounding rockets proceeded at a furious pace in the years immediately after World War II. Between 1946 and 1950, the United States flew almost 100 V-2 and Aerobee rockets from the White Sands Missile Range with instruments that measured cosmic rays, solar radiation, atmospheric characteristics, and sky brightness. Cameras were flown that photographed Earth; even biological experiments were carried out. During this period, scientists were learning how to use rockets. The two outstanding problems that had to be dealt with were controlling the orientation of the rocket and retrieving the data and the instrument.

To achieve stable flight, rockets are generally spin stabilized. The rocket fins are tilted slightly so that aerodynamic forces induce a spin in the rocket as it ascends through the atmosphere. The effect is the same as the “rifling” in a gun barrel that imparts spin to a speeding bullet. Thus a rocket will maintain the same heading, more or less, during its flight as it had when it began at the ground. For most of us these days, the image we have of a rocket launch is NASA’s space shuttle. Such large rockets begin their flight too slowly to achieve much spin. They are held steady by gyroscopes that determine the orientation

and small vernier rockets that keep the rocket headed in a set direction. For the earliest rockets observations, the rocket was pointed at a place such that a given object or region was in view during all or part of the flight taking into account the spin of the rocket. For some experiments, for example, measuring cosmic rays, pointing up was sufficient; for others, such as solar observations, lack of control limited the kinds of observations that could be conducted. The V-2 did use gyroscopic control because it was intended to impact a specific point and needed good control during the early phase of its flight.

Retrieving data by using radio communications to the ground was straightforward for data that could be in an electronic form. However, much of what scientists wanted to do, especially imagery, could not be reduced to electronic signals. So it was necessary to devise a means to recover parts of the rocket, at a minimum. Initially, the approach was brute force. The sensing portion of the payload would be built into a "crash-proof" container. Following some well-publicized instances where the rocket and its scientific payload produced a large crater and little of recoverable value upon its return to Earth, the V-2s were equipped with explosive charges that separated the science payload from the rocket, and the payload would then flutter down to Earth at lower speed. Eventually, Aerobees were equipped with parachutes that allowed recovery of the payload, more or less intact. Recovery of the payload also had the advantage of reducing the cost of reflights of the instrument.

A cosmic-ray experiment has the honor of being the first experiment performed from a sounding rocket. In April 1946 and twice again in May, single Geiger counters were flown by James Van Allen and his group at the Applied Physics Laboratory on V-2 rockets (11); in the first few years of rocket science, cosmic-ray measurements were the most popular experiments to be conducted. These experiments achieved a fair degree of sophistication. The first NRL experiment, flown in June 1946, used 10 Geiger counters, interspersed with 10-cm lead shielding, that weighed a total of 100 pounds (12). A number of important results emerged during the first several years of sounding rocket work. Van Allen and Singer (13) demonstrated that cosmic-ray intensity depends strongly on the strength of the local magnetic field. They reported that the intensity at the magnetic equator was about a factor of 3 lower than at White Sands, based on the flight of an Aerobee rocket from a U.S. naval vessel. The higher magnetic field at the equator sweeps away lower energy particles that can penetrate nearly to the surface of Earth at New Mexico. Going further north where the field is weaker, cosmic ray intensity increases further. Aside from measuring its intensity, scientists measured its atomic composition. Cosmic rays are mostly protons, but other heavier constituents were known to be present. In a 1950 V-2 rocket flight, the NRL group, using a detector assembly that had 48 Geiger counters, 2 ionization chambers, and about 8-inch thick lead slabs measured the presence of helium nuclei in cosmic rays and possible other elements as well, aside from the dominant proton contribution (14). Figure 2 illustrates the data obtained from this kind of experiment. These data came from X-ray detectors flown in 1962 that detect cosmic rays as well. On the ground, the counting rate is very low. As the rocket lifts off, the counting rate rises rapidly, reaches a peak at 52,000 ft (16 km), then falls to a steady value between about 117,000 and 250,000 ft. The peak is the Pfofzer maximum. It results because every cosmic ray that enters the

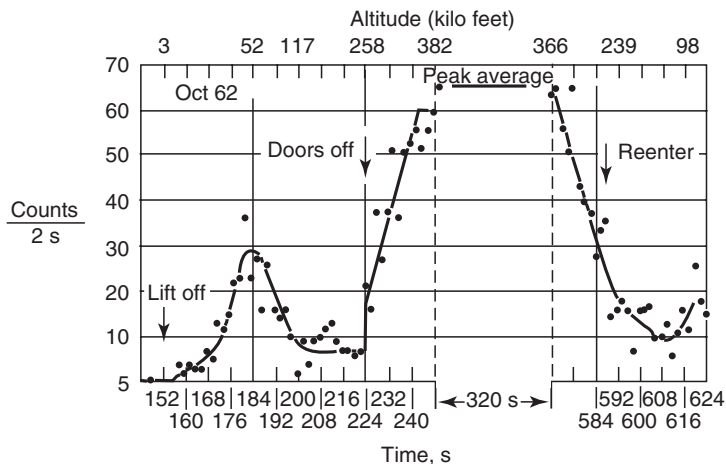


Figure 2. Data from a sounding rocket flight showing the changes in counting rate in radiation detectors as the rocket ascended through the atmosphere and returned.

atmosphere produces many particles when it interacts, which in turn produce more particles when they interact. This process continues until the average energy of the particles degrades, and multiple particles are no longer produced. So the number of particles increases and then decreases. At 250 seconds, the detectors, which are in the rocket and are shielded by a heavy aluminum door, are exposed to the outside environment and can respond to much lower energy radiation than represented by cosmic rays. The counting rate jumps abruptly and continues to rise as the rocket leaves the last vestige of the atmosphere. The process is repeated 300 seconds later as the rocket reenters.

Cosmic-ray experiments, aside from being first, also illustrate the limitations of sounding rockets, namely, time at altitude and payload weight. The cosmic-ray intensity is very low, only about one per cm^2 per second above the United States. They are also highly penetrating and require the use of lead or other heavy shielding as part of the experimental apparatus. The net result was that the cosmic-ray community had to await the development of large boosters and orbiting spacecraft before they could make significant advances in their discipline. Even today, cosmic-ray experiments strain the capability of the various national space agencies with their size and weight requirements.

Early solar observations were hampered by the fact that film was being used as the recording medium and needed to be recovered after the flight. Both NRL and APL developed spectrographs to measure the ultraviolet spectrum of the Sun, a subject of intense interest to astronomers of the day. The approach at that time was to put the film into an armored vessel, which eventually proved successful. Another problem facing the experimenters was that there was no good way to point continuously at the Sun. As a solution, both the NRL and the APL experimenters chose to place spectrograph apertures on opposite sides of the rocket and took short exposures to catch the Sun just at the moment when it shone in these openings, as the rocket spun on its axis. Tousey and his group at NRL were the first to obtain UV spectra during a V-2 flight on 10 October 1946

(15). During the next several years, both NRL and APL obtained successful spectra to a limiting short wavelength of about 2200 Å that revealed hundreds of spectral lines. However, these results, which were important for the solar physics community, did not address a major issue of interest to NRL, namely, what was the radiation responsible for the ionosphere. To answer that question, it was necessary to search for X rays. The first attempt to do so by T.R. Burnight of NRL during an Aerobee rocket flight of 5 August 1948 used X-ray sensitive phosphors and film. As all too often occurred during these early days of rocket science, successful detection was claimed, but the result was ambiguous (16). Subsequently, other NRL scientists used thermoluminescent crystals to attempt to detect X rays. These are crystals that are excited by X radiation and subsequently glow when heated in proportion to the initial excitation. The issue was resolved decisively by Herbert Friedman and his colleagues at NRL, during a V-2 rocket flight of 29 September 1949 by using X-ray sensitive Geiger counters that could record individual X-ray photons (17). Friedman was a newcomer to atmospheric and space research. His background was laboratory X-ray analysis, and he had spent the war years applying his specialty to military needs. During that time, he had developed Geiger counters with very thin walls that were sensitive to X rays.

The requirements of the solar researchers led to the first important advance in rocket science since the emergence of the V-2 rockets themselves, namely, the development of a means to point the payload at a given target. In 1947, both NRL and APL built devices that tracked the Sun. About the same time, Marcus O'Day of the University of Michigan provided Air Force funding to a group at the University of Colorado to develop such a device. The Colorado device, a two-axis system, flew first in 1949 and became the standard unit for studies of the Sun. It also led to one of the most successful commercial "spin-offs" from rocket science. A few of the engineers from the University of Colorado formed a private company to provide support for the development of experimental rocket payloads, supported by financing from the Ball Brothers Corporation of Muncie, Indiana. The company is now Ball Aerospace Corporation.

Studies of the ionosphere were equally popular. Unlike cosmic rays, which could be measured in the atmosphere, the ionosphere begins at an altitude of about 100 km. Thus rocket-borne instruments provided the first direct reach into this atmospheric region. Most of the measurements, however, were not direct; rather, they relied on measuring the effects on the propagation of radio waves from the rocket to a ground receiver. However, these yielded detailed information on the region's structure. A direct measurement of electron density was made by the University of Michigan group from a December 1947 V-2 rocket flight using a Langmuir probe, an instrument invented in the 1920s to measure the electrons produced in an electrical discharge. In this case, the ionosphere itself represented the discharge, and the electrons comprising the ionosphere were collected as a current. This instrument, in various forms, has become a staple of space research to record ambient electrons and their accompanying protons.

Studies of the ionosphere and companion measurements of the neutral atmosphere verified and extended the general understanding of the upper atmosphere. However, the limitation of rocket science here also became important. In contrast with cosmic-ray science, the problem was not weight limitation.

Instruments for measuring the atmosphere were small then and still are; rather, the problem was that a single rocket flight provides only a snapshot of a complex phenomenon that varies in space and time. Again, only the advent of orbiting spacecraft made it possible to extend measurements over the whole earth and through many seasons and allowed developing a thorough understanding of the underlying phenomena. The same is true of most geophysical phenomena.

An important area of applied research that attracted the attention of the early rocket scientists was photography of Earth from space. In 1947, groups from both APL and NRL were successful in obtaining photographs from V-2 rockets. Figure 3 shows a typical photograph from a rocket obtained during a V-2 flight in 1947. But because rocket cameras could take pictures for only a few moments in a restricted place, their importance as a practical reconnaissance or meteorological tool was recognized as limited in the same way as were other geophysical investigations. NRL did continue with a program of rocket photography, principally to determine where the rocket was pointing. However, in 1954, the group fortunately caught a hurricane in its view. As explained by the NRL researchers, rockets were seldom fired if the cloud cover exceeded 10% because that would make it difficult to photograph the rocket from the ground and to obtain photos of ground features from the rocket. However, they go on to note,

“A fortunate exception occurred on October 5 1954. Two rocket-borne movie cameras obtained pictures of towering clouds spiraling into a tropical storm near Del Rio, Texas.”

The resulting composite photograph, covering more than a million square miles of area clearly showed the utility of photographing storm clouds from the vantage point of a rocket (18).



Figure 3. Photograph of Earth taken in 1947 at an altitude of about 160 km from a V-2 rocket flown from the White Sands Missile Range.

In spite of the emergence of important new results, some university groups (Princeton, Harvard, Cal Tech) believed that the prospects were so limited that they abandoned the idea of using rockets for science projects. However, what was demonstrated was more important than any science. The several active groups showed that it did not take heroic efforts to build scientific payloads. These groups used instruments adapted from their laboratories or entirely novel instruments and flew them repeatedly on short schedules. Some of the basic requirements for rocket scientists were solved: notably retrieving data from space using radio technology and recovering payloads. A start was made in developing actively controlled rockets that could point at specific targets. The United States developed new rockets and improved their reliability enormously. By the 1950s, rockets, more often than not, did work, in contrast to the first few years, when the reverse was true. These advances had the effect of making rocket science more attractive to individual researchers and to funding agencies.

A Maturing Discipline

By the mid-1950s, many of the developments in contemporary rockets were in place. The Aerobee was replaced with the Aerobee-Hi, a rocket developed by the U.S. Air Force exclusively for sounding rocket research and designed to carry a parachute for payload recovery. The concept of controlling rockets had emerged with the successful development of the Sun seeker. A major scientific conference exclusively dedicated to rocket science was held in 1952 in Oxford, England, sponsored by the Rocket Research Panel of the United States and the Royal Society of London (19).

In 1957, events occurred that changed space science. The Soviet Union launched the world's first artificial satellite in October of that year, as part of its contribution to the International Geophysical Year. The complementary U.S. effort was successful in the following year. The subsequent emergence of a national space agency (NASA) in the United States and the strong support of space science increased significantly the available funding and the number of participating groups. Other nations had also made commitments to national programs of space science. The focus of these programs was orbiting satellites, but the new interest did result in a significant increase in the number and quality of rockets available for research. There are a few new rockets, notably, the Canadian Black Brant, that replaced the Aerobee, and there has been considerable development of the means to control the rocket itself. In addition to high precision "Sun followers," "star followers" are now available that can maintain subarc minute pointing precision.

There have been surprises, one of which was in the X-ray domain. The observation of X rays from the Sun had been one of the great successes of early rocket science, both from the perspective of the novel detectors that had been developed and the richness of the phenomena that had been observed. Thus, it was natural to consider looking for X rays from cosmic sources. The likelihood of observing Sun-like stars was extremely remote. Even the nearest stars are about a million times more remote from us than the Sun, and their radiation is reduced by 10 trillion times compared to that of the Sun. Nevertheless NRL's Herbert

Friedman did attempt to observe cosmic X rays from rockets without success. In 1960, he wrote,

Efforts to observe X-ray emission from celestial sources have yielded only negative results. On the basis of three experiments thus far, it can be stated that no fluxes stronger than ... the 1–10 Å and 44–80 Å bands have been seen in fairly complete scans of the sky from White Sands, New Mexico. The observation of smaller fluxes will require longer observing times, available from satellite platforms rather than rocket probes. (20)

Yet within a few years, X-ray astronomy from sounding rockets became an exciting new field of science.

In 1962, a group of scientists from American Science and Engineering, a private company in Cambridge, Massachusetts, flew an Aerobee sounding rocket from the White Sands Missile Range with Air Force sponsorship. The principal objective of the rocket was to look for X rays from the Moon that should result from excitation by solar X rays. A further objective was to survey the sky for celestial X-ray sources. The result was the detection of a very strong source of radiation near the galactic center and an indication of other sources as well (21). The source is now known to be Scorpius X-1, still the brightest steady X-ray source in the sky. This result was confirmed and extended, notably, by Friedman at NRL (21). Within a few years, X-ray observations of the cosmic sources became recognized as a distinct part of astronomy on a par with radio astronomy and cosmic rays. X-ray astronomy is an example of a discipline that emerged only because of the availability of rockets. Though X-ray observations are still carried out by rockets, satellite observations have dominated the discipline since 1970 when the first satellite dedicated to X-ray observations, UHURU, was launched into orbit.

The discovery of bright X-ray sources in 1962 marked another kind of transition. After it was established in 1947 as a service independent of the U.S. Army, the Air Force vigorously pursued a space mission, including sponsorship of basic scientific research from sounding rockets and satellites. Secondly, the research group at American Science and Engineering, Riccardo Giacconi, Frank Paolini, and the author, were in secondary school when the war ended and the first rockets were being flown for research purposes. This was a second generation from the perspective of sponsorship and personnel. The 1960s also marked the decade that NASA, founded in 1958, was creating a new constituency for space sciences. By 1964, NASA was committing substantial resources to the new discipline of X-ray astronomy.

Sounding rockets still fill an important niche in space sciences, especially in demonstrating new kinds of observational capability. NASA launched 20 rockets in 1999, 10 from the White Sands Missile Range. The image shown in Fig. 4 is an example of the sophistication that can be achieved from a rocket. It is an image of the Sun's atmosphere in the radiation from hot hydrogen. This radiation emerges from the Sun's chromosphere, immediately above the Sun's surface. The spatial resolution of the image is about one-third of an arc second and is a tribute to the instrument developers at the Naval Research Laboratory and to the rocket developers at NASA. Results of this sort are still an important source of scientific

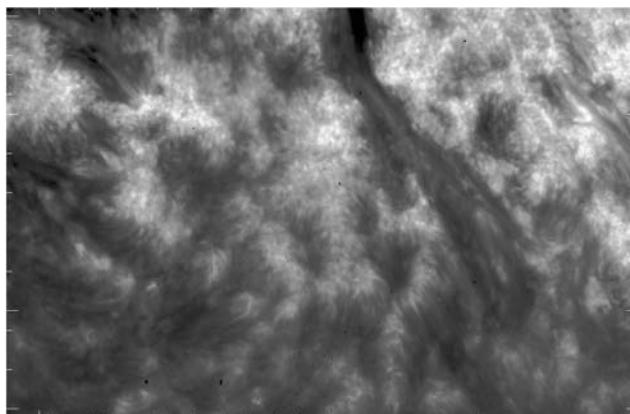


Figure 4. Image of a portion of the Sun's lower atmosphere taken in the light of very hot hydrogen. The hydrogen is confined to the Sun's magnetic field, yielding the strand-like appearance. The many dark features reveal places where the Sun's magnetic field emerges from below the surface. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

results of high importance and a guide as to what is likely to make a productive space mission. Their cost is now well above the hundred thousand dollars or so that rocket experiments used to cost but still well below the hundred millions of dollars or so of satellite programs.

BIBLIOGRAPHY

1. Newton, I. *Opticks*. Dover, New York, 1952 (originally published in 1703).
2. New York Times, *Topics of the Times*, 13 January 1920.
3. Winter, F.R. *Rockets into Space*. Harvard University Press, Cambridge, MA, 1990.
4. Piccard, A. *Literary Digest*, Jan. 28, 1933.
5. Labitzke, K.G., and H. van Loon. *The Stratosphere: Phenomena, History and Relevance*. Springer-Verlag, New York, 1999.
6. Encyclopedia Britannica (1902), quoted by Hevley, B.W., Basic Research Within a Military Context: The Naval Research Laboratory and the Foundations of Extreme Ultraviolet and X-ray Astronomy, 1923–1960, University Microfilms International, 1992.
7. Rossi, B. *Cosmic Rays*. McGraw-Hill, New York, 1964.
8. Dornberger, W. V-2. Viking, New York, 1954.
9. DeVorkin, D.H. *Science With A Vengeance*. Springer-Verlag, New York, 1992.
10. Chodil, G., H. Mark, R. Rodrigues, F. Seward, and C.D. Swift. *Rev. Sci. Inst.* 38: 11508 (1967).
11. Van Allen, J.A., and H.E. Tatel. *Phys. Rev.* 73: 245 (1947).
12. Golian, S.E., et al. *Phys. Rev.* 70: 776 (1946).
13. Van Allen, J.A., and S.F. Singer. *Phys. Rev.* 78: 819 (1950).
14. Perlow, G.J., et al. *Phys. Rev.* 88: 321 (1952).
15. Johnson, F.S., J.D. Purcell, and R. Tousey. *J. Geophys. Res.* 56: 583 (1951).
16. Burnight, T.R. *Phys. Rev.* 75: 165 (Abstract) (1949).
17. Friedman, H., S.W. Lichtman, and E.T. Byram. *Phys. Rev.* 83: 102 (1951).

18. Hubert, L.F., and O. Berg. *Mon. Weather Rev.* 119: June 1955. (An image of the storm reported in this paper also appeared in *Life Magazine*, 5 September 1955.)
19. Boyd, R.L.F., and M.J. Seaton (eds). *Rocket Exploration of the Upper Atmosphere*. Pergamon, London, 1954.
20. Friedman, H. *Proc. 10th Int. Astrophys. Symp.* Liege, Belgium, 1960.
21. Giacconi, R., H. Gursky, F.R. Paolini, and B.B. Rossi. *Phys. Rev. Lett.* 9: 439 (1962).
22. Bowyer, C.S., E.T. Byram, T.A. Chubb, and H. Friedman. *Nature* 201: 1307 (1964).

HERBERT GURSKY
Naval Research Laboratory
Washington, DC

SIZE AND SHAPE OF EARTH FROM SATELLITES

Modern evidence abounds that clearly supports the idea that the Earth is a spherical body. The memorable photographs of the Earth taken by the Apollo astronauts from the lunar surface and from cislunar space must motivate serious consideration of the sphericity notion even among the hardest skeptic.

But the stirring evidence presented in a photograph from space was unavailable to the ancients who had to deduce the ideas of sphericity from subtle effects (such as the sails of a ship slowly sinking at the horizon), from obvious comparative forms (the Moon and the Sun appear to be spheres), and from philosophical ideas (the sphere is a perfect geometric object, hence the Earth must be a sphere). Although these early approaches spurred the development of the science of geodesy, the revolutionary work of Isaac Newton in the seventeenth century brought the first renaissance in geodesy. The second renaissance coincides with the space age, which brought new tools to the pursuit of ancient questions. The new discipline of space geodesy appeared in the early 1960s, but the areas of geodetic interest go far beyond the classical study that focused on the size and shape of the Earth. Modern geodesy includes scholarly pursuits into the mass distribution within the Earth, as manifested by the external gravity field, the rotation of the Earth, and temporal changes in the positions of points on the Earth's surface. The modern tools and techniques of geodesy are used by other scientific disciplines, including tectonophysics, oceanography, and glaciology, to name a few. And the methodologies and concepts are applied to other planets in the discipline known as planetary geodesy.

Historical View

The notion of a spherical the Earth is often attributed to Pythagoras in the sixth century B.C., but Eratosthenes made the first serious attempt to measure the geometric characteristics of such a model in the third century B.C. As the

librarian at Alexandria, Egypt, it is thought that he was one of the most learned men of antiquity (1). He made the subtle observation that the Sun cast a shadow of different lengths from an upright rod at two different locations, even though observed at the same time of year.

Eratosthenes must have observed the changes in the length of the shadow during the course of a year near his Alexandria home. But the shadow cast under similar conditions near today's Aswan Dam was noticeably different. Assuming that the Sun's rays arrived parallel at both Alexandria and Aswan, coupled with knowledge of the distance between the two sites (about 800 km), Eratosthenes computed an approximate radius of the Earth. This was a remarkable application of known geometric properties of the day, but assessing the accuracy of the radius by comparison with a modern estimate of the Earth's radius depends on the conversion between distance units that Eratosthenes used and those used in modern geodesy. Although the precise conversion is debated by scholars of science history, the worst case suggests that Eratosthenes was in error by 15%, a rather remarkable achievement in the third century B.C. Eratosthenes is sometimes referred to as the father of geodesy, a well-justified title.

In the second century B.C., the astronomer Hipparchus determined a subtle motion of the Earth that remained unexplained for 1500 years (1). He noted that star positions had shifted systematically since the determinations by others 150 years earlier. This shift, now known as the precession of the equinoxes, amounts to only $1.4^\circ/\text{century}$. The explanation of this motion of the Earth had to wait for the brilliant work of Isaac Newton, who calculated the rate produced by lunisolar forces in response to an interaction with the Earth's mass distribution and, hence, its shape. Other more subtle motions of the Earth, analogous to those of a gyroscope, which include nutations, also result from the interaction of the Sun and Moon with the mass distribution of the Earth (2).

The First Renaissance. There is little documentation of efforts to measure the radius of the Earth in the 1500 years after Eratosthenes. Nevertheless, the voyages of exploration in the fifteenth and sixteenth centuries were clear demonstrations of the spherical nature of the Earth.

In the seventeenth century, geodesy was invigorated by several technological and scientific developments. These developments included technologies for measuring the length of a specified latitudinal displacement, such as 1° . Using the technology to measure the difference in latitude between two points, a second technology was needed for geodetic applications to measure the distance between these points. The latitude measurements were made with astronomical instruments, known as transit circles, and early versions of angle measuring surveying equipment were developed to support the distance measurement.

In the early 1600s, Willebrord Snellius (or Snell) published a new method that would significantly contribute to determining the shape of the Earth. This method allowed accurate inference of the distance between two points at different latitudes along the same meridian of longitude. This innovative method diminished the effect of local topographic obstacles that were inherent in the direct measurement of distance between two points separated by 100 km (1° of latitude) or more. In essence, Snell's method made an accurate measure of the distance between two points on flat terrain, usually separated by a kilometer or two. This baseline measurement was made by placing a series of calibrated rods end to

end. Using a series of triangles and the measurement of all angles within the triangles, the single baseline distance could be used to determine the lengths of all triangle sides, regardless of the intervening terrain. Another baseline measurement was usually made with the rods elsewhere in the network to verify the result from the triangles. This method of triangulation was the basic methodology for determining the lengths of various meridian arcs that were made during two centuries.

Using the new methods, an important set of arcs was measured in France during the late 1600s and early 1700s. The measurements made by the Cassini family within France tended to support the view that the Earth was not a perfect sphere; instead, the results suggested that the Earth was somewhat elongated along the polar axis and squeezed inward at the equator. A geometrical figure that matched the Cassini results was a prolate spheroid, which can be created by an ellipse rotated about its major axis. In this model, the Earth's polar axis would coincide with the major axis of the ellipse.

At the same time, Newton published his monumental *Principia* in 1687 (3). In this work, Newton showed that the centrifugal force resulting from rotation of a sphere would tend to distort the sphere around the equator, so that the body would be characterized as an oblate spheroid. This body is generated by rotating an ellipse about its minor axis; and the minor axis coincides with the polar axis.

Proponents and detractors were attracted to the Cassini view that the Earth was a prolate spheroid, whereas others supported the Newtonian view of an oblate spheroid. The scientific debate occasionally degraded into nationalistic fervor to support one or the other view. Because of heated interest in the topic, the Royal Academy of France in 1735 suggested that expeditions be undertaken to measure the length of arc meridians at different latitudes. The first expedition was directed to take place in South America near the equator (modern Ecuador), creating what became known as the Peruvian arc. A second expedition was dispatched to Lapland, in polar Scandinavia. If the Earth was an oblate spheroid, then 1° of latitude will be slightly longer in the polar region than in the equatorial region. Information about the expeditions has been summarized by Todhunter (4) and Butterfield (5).

Both expeditions experienced great hardships. The Lapland expedition left a year after the Peruvian group, but it was completed within a year, thereby giving it the honor of publishing the first results in support of the Newtonian view. The Peruvian expedition endured many difficulties, which were compounded by rugged terrain and infighting among the participants. It took 10 years for the expedition to return to Paris and make known their results. But the results again reinforced the Newtonian oblate spheroidal model of the Earth. The philosopher Voltaire made the observation about the expeditions: "You have found by prolonged toil, what Newton had found without even leaving his home" (6).

The oblate spheroid model of the Earth is characterized by (1) the equatorial radius (the ellipse semimajor axis, but it is a circle when rotated about the polar axis); and (2) the polar axis (the ellipse semiminor axis). Instead of using the specific polar dimension, an alternate parameter known as flattening is used, which is related to the eccentricity of the ellipse. If a represents the semimajor axis and b is the semiminor axis, then flattening f is $(a - b)/a$.

Although the Peruvian and Lapland expeditions showed that the Earth was more appropriately represented by an oblate rather than prolate geometric figure, the equatorial radius and flattening parameters that resulted from the expeditions were significantly different, ranging from $1/178$ to $1/266$ for f . In the following 200 years, geodesists whose names became associated with them made various determinations of the ellipsoid parameters: Airy, Everest, Clarke, Helmert, and Hayford to name a few (6).

Perhaps the most innovative presatellite determination was made by Sir Harold Jeffreys (7). He used the measurements of arc lengths, and he also included measurements based on the changes in pendulum motions as a function of latitude and the precession of the equinoxes. His combination solution gave 297.1 for $1/f$.

The Shape and Gravity Field of the Earth

Newton's law of gravity states that the magnitude of the gravitational force F between two particles separated by a distance d is GM_1M_2/d^2 , where G is the constant of gravitation and M_1 and M_2 are the masses of the particles. The mathematical model for the particle is a point mass, that is, it is assumed that all of the mass of the body is concentrated at a point. Gravity is an attractive force, so that each mass experiences a force of the same magnitude, but the force on M_2 is directed toward M_1 , for example. The acceleration of gravity g experienced by M_2 as a falling object is given by GM_1/d^2 . If M_1 represents the Earth and d is the distance between the center of the Earth (assumed as the location of the point mass model), a falling body M_2 will experience an acceleration of 9.8 m/s^2 near the surface of Earth.

The point mass concept applied to the Earth is a serious stretch of the imagination when one views the Earth as a resident on the surface. Nevertheless, it can be shown that a spherical body that has a uniform mass density is gravitationally equivalent to the point mass concept. But if the body is an oblate spheroid, even with uniform density, the force experienced by a point mass external to the spheroid will not correspond exactly to the simple inverse square of the distance law. This statement does not mean that Newton's law of gravity is invalid.

Consider that the oblate spheroid figure is subdivided into a very large number of point masses or a set of finite mass elements. Each point mass interacts with an external mass, M , in accordance with Newton's law of gravity. But the total force experienced by M is the result of the summation of the individual finite element contributions. This net force will be different from the force of gravity computed from the total mass of the spheroid, the distance from the spheroid's center, and the mass M . The source of the difference is caused by the fact that a point mass mathematical model is not a satisfactory representation of an oblate spheroid or a general geometric body that has an arbitrary distribution of interior mass, but the model does apply to the interaction between the external mass and the finite mass elements. It can be shown that a scalar potential function U can be used to describe the gravitational field of a body that results from any mass distribution within the body. The usual

expression for U is (8)

$$U(r, \phi, \lambda) = \frac{GM_E}{r} + \frac{GM_E}{r} \sum_{l=2}^{\infty} \sum_{m=0}^l \left(\frac{a_e}{r}\right)^l P_{lm}(\sin \phi) (C_{lm} \cos m\lambda + S_{lm} \sin m\lambda), \quad (1)$$

where

GM_E = the gravitational parameter of the Earth

a_e = the mean equatorial radius of the Earth

r, ϕ, λ = the radial distance, the geocentric latitude, and the longitude of an external point

$P_{lm}(\sin \phi)$ = the associated Legendre function of degree l and order m

C_{lm}, S_{lm} = the spherical harmonic coefficients of degree l and order m

When $m = 0$, an alternate set of coefficients is often used such that $J_1 = -C_{1,0}$. These coefficients are referred to as zonal harmonics because the representation divides the gravitational effect into zones of latitude. When $l = m$, the coefficients are known as sectoral harmonics because the gravity field is divided into sectors of longitude. Finally, all other coefficients are referred to as tesseral harmonics. It is important to note that zonal harmonics are associated with terms that have no longitudinal dependence. For an oblate spheroid of uniform density, the gravitational potential requires only terms of even degree l . Using this scalar function, the force of gravity that a point mass M will experience is

$$F = M \nabla U. \quad (2)$$

A fundamental question now arises: What is meant by the shape of the Earth? On the one hand, it is immediately evident that if detailed account is taken of all of the topographic surface variations (mountains, valleys, ravines, human landforms, and structures), even at the scale of 1-kilometer, there is no simple geometric figure analogous to the oblate spheroid. An alternative description is based on a shape that reflects the mass distribution within the Earth, that is, the gravity field. As it turns out, a simple geometric figure cannot be assigned to either description unless one is not particularly concerned how accurately well the figure represents the actual Earth.

The shape of the Earth defined by the gravity field uses the potential function U . If a surface of constant potential is defined such that $U = \text{const}$ everywhere on the surface, the shape of this equipotential surface will serve as a surrogate shape of the physical Earth. The selection of the constant value remains, but if the constant is chosen so that the surface coincides with mean sea level, the surface has a readily understood relevance to everyday experiences. This particular surface is known as the geoid, but the gravitational potential given before must be augmented by a term that will account for the centrifugal force due to the Earth's rotation. The definition of the geoid is based on an augmented potential W given by $U + \frac{1}{2} \omega^2 r^2 \cos^2 \phi$, such that $W = \text{const} = W_0$, and ω is the rate of the Earth's rotation based on the sidereal day ($2\pi/86164$ rad/s). The gradient of W , multiplied by the mass of a body at rest on the rotating Earth,

yields the weight of the body, which includes the force of gravity as well as the centrifugal force caused by the Earth's rotation.

The gravitational field of an oblate spheroid, an ellipse of revolution that has uniform density, is represented by the potential function

$$U = \frac{GM_E}{r} - \frac{GM_E}{r} \sum_{l=2}^{\infty} \left(\frac{a_e}{r}\right)^l P_{l,0}(\sin \varphi) J_l, \quad (3)$$

where l is an even number. To a reasonable level of approximation, the flattening of the ellipsoid representing the geoid is related to the second-degree zonal harmonic J_2 by the relationship

$$f = \frac{1}{2}(3J_2 + K) \left(1 + \frac{3}{4}J_2 + \frac{3}{28}K\right), \quad (4)$$

where K is related to the ratio of centrifugal to gravitational acceleration at the equator, 0.0034498 (9). Although the representation has little practical use in modern geodesy, it illustrates the conceptual relationship between the parameters.

The Space Age

The First Decade. The second renaissance in geodesy began with the launches of the first artificial satellites in the late 1950s. Whereas the determination of the Earth's shape in the eighteenth century was based on measurements of meridian arcs, the opportunity presented by artificial satellites required understanding the influence of the Earth's gravity on the satellite orbits. A fundamental question is: How will the orbit of a satellite respond to the mass distribution of an oblate spheroid?

The orbit of a satellite is characterized by six orbital elements: three that describe the orbit geometry and a reference time, plus three that describe the angular orientation of the orbit in space. The former three are semimajor axis, eccentricity, and a reference time at perigee passage. The spatial orientation of the orbit is defined by the location of the ascending node, the inclination of the orbit plane with respect to the equator, and the angular location of perigee (10). The ascending node Ω is an angle measured eastward at the equator from a fixed direction, usually the vernal equinox, to the point where the satellite crosses the equator from the Southern Hemisphere into the Northern Hemisphere. This angle is usually referred to as the right ascension of the ascending node. If the Earth were a perfect sphere of constant density (gravitational equivalent of a point mass) and the satellite were influenced only by the Earth's gravity, then the six orbital elements for the satellite orbit would be constant, and the specific elements depend on the position and velocity of the satellite at some time. This is the major result of the classical problem of two bodies.

The mass distribution of an oblate spheroid, for example, perturbs the orbit from the ideal two-body motion. The consequence of this perturbation is that the six orbital elements are functions of time. Time variations in the orbital elements

are usually characterized by (1) linear variation in the average value over time; and (2) periodic variations, which are usually dominated by a cyclic change that goes through two cycles in one orbital revolution (a twice per revolution effect). It can be shown that the linear variation, known as secular change, depends on all even-degree zonal harmonics, but in the case of the Earth, the second degree zonal harmonic is of order 10^{-3} , which is 1000 times larger than any of the other zonal harmonics. Consequently, this degree-two term dominates the description of secular motion. Only the location of the ascending node and perigee exhibit the secular changes that are of interest here. It can be shown that the secular temporal rate of change Ω in the right ascension of the ascending node of a satellite orbit is

$$\frac{d\Omega}{dt} = -\frac{3}{2}J_2 \frac{n}{(1-e^2)^2} \left(\frac{a_e}{a}\right)^2 \cos i, \quad (5)$$

where the satellite orbit parameters are the mean motion n , semimajor axis a , eccentricity e , and inclination i . The value of J_2 will be positive for an oblate spheroid and negative for a prolate spheroid. Consideration of the node rate equation shows that the ascending node will regress for an oblate spheroid and posigrade inclination. In other words, the ascending node will move in a westward direction with respect to the stars.

If the node rate of a satellite is observed and the orbital elements (a, e, i) are known, then the gravity coefficient J_2 can be determined from Equation 5. Using J_2 , the flattening f can be found from Equation 4 that relates f and J_2 . This is an oversimplified illustration of how the Earth's shape can be determined from observations of a satellite's motion.

The launch of Sputnik I by the Soviet Union on 4 October 1957 opened the space age. The satellite orbit decayed because of atmospheric drag and a low perigee and the satellite was destroyed in atmospheric reentry on 4 January 1958. The 83-kg satellite operated on batteries and transmitted temperature data for 23 days in orbit. Even though the orbit of the satellite was not well determined, the westward motion of Ω was evident, and the satellite provided an early assessment of the prelaunch parameters that described the shape and mass distribution of the Earth.

One month after Sputnik I, the much larger Sputnik II (507 kg) was launched in November 1957, carrying a dog. The solar-illuminated satellite was photographed against the stars, thereby enabling a reasonably accurate determination of the orbit. Early comparisons by Merson and King-Hele (11) of the predicted node rate (based on the best available flattening) with the observed rate suggested a discrepancy in flattening of about 1%, compared to the presatellite value.

Although the United States was stunned by the Soviet achievements, work had already been underway to launch a satellite for U.S. participation in the International Geophysical Year. As early as 1954, Major John O'Keefe of the Army Map Service suggested that a satellite would allow studying the size and shape of the Earth and the intensity of its gravitational field (12).

In January 1958, the United States launched Explorer I, followed by Vanguard I on 17 March of the same year. The Vanguard Project was under intense

scrutiny because of early launch failures, but the project had been well planned. A network of tracking stations, known as Minitrack, had been deployed that was ready for postlaunch operations, and the use of a state-of-the-art IBM 704 digital computer had been arranged to determine the orbit. The instrumentation carried on the 2-kg satellite included batteries, solar cells, radio transmitters, and temperature sensors. The satellite represented a significant achievement in miniaturization, and the Minitrack system proved capable of providing accurate measurements of angles for determining the orbit by using a radio interferometric technique. Nine Minitrack stations were distributed across North and South America, plus a station in Australia and one in South Africa. The objectives of Vanguard I were to “determine atmospheric density and the shape of the Earth, to evaluate satellite thermal design parameters and to check the life of solar cells in orbit” (13).

In February 1959, using Vanguard I data, O’Keefe (by then associated with NASA), published the surprising result that suggested the existence of J_3 , which would relate to a pear shape of the Earth. In September 1959, he published values for the zonal harmonics up to degree four, the first such determinations for the third- and fourth- degree zonal coefficients (14,15). The key to determining J_3 was the fact that odd-degree zonal harmonics produce asymmetry in a geoid with respect to the equator. Whereas the oblateness of the Earth is characterized by symmetry with respect to the equator, which produces a secular, or linear, change in the location of the ascending node and perigee location of the orbit, the effect of asymmetry is quite different. Examination of the eccentricity variation in the Vanguard I orbit revealed a long-period change of 82 days, the interval required for perigee to make one revolution in response to the oblateness, or J_2 . O’Keefe estimated that the pear shape corresponded to a 15-meter undulation of the geoid. This undulation compares to the dominant characteristic of J_2 , which is associated with the fact that the difference between the equatorial and polar axes is about 21 km.

The objectives of Vanguard I were achieved. The pear shape contribution to the shape of the Earth was hailed as one of the major, and perhaps unexpected, results of Vanguard. The refinement of a model for the figure of the Earth continued through the early 1960s, but the studies were limited by a paucity of data. It was fortunate that Vanguard I had been designed with solar cells because the battery failed after a few months in orbit. But Vanguard continued transmitting for several years, thereby enabling observations on the nature of its long-term orbit evolution. It has been predicted that Vanguard I, now quiet, will remain in orbit until well into the twenty-third century or longer.

Thus began the discipline known as *Satellite Geodesy*. Project Vanguard foretold the requirements for continued improvements in satellite geodesy: satellite instrumentation interacting with a global network of tracking stations to provide accurate observations related to satellite position and digital computers to analyze the observations. In April 1962, the first international symposium on the use of artificial satellites for geodesy was convened in Washington, D.C. (16). The rapidly expanding discipline attracted presentations by almost 50 individuals and gave strong evidence of the synergy with celestial mechanics, geophysics, satellite tracking systems, reference frames, and numerical analysis. The publication by Kaula (17) of a gravity field for the Earth to degree and order eight

based on Sputnik II and Vanguard I was an early indicator of the contributions that satellites would make in determining the shape of the Earth as reflected by the geoid. The Earth's parameters determined by Kaula were equatorial radius = 6378163 m and $1/f = 298.24$.

The oblate shape of the Earth and the pear shape produce distinctly different effects on the motion of an artificial satellite. But other unusual effects exist. Even though the first geosynchronous satellite (an orbital period equal to the Earth's rotational period) did not begin operation until July 1963 (Syncom 2), the special interaction between the ellipticity of the Earth's equator and satellite motion was highlighted at the 1962 symposium. If the equator was elliptical, there would be only four equilibrium locations where a geosynchronous satellite would remain stationary (the points are on opposite sides along the major and minor axes), although the problem of two bodies predicted an infinite number. As it turned out, the two points along the minor axes are dynamically stable, which implies that the geosynchronous satellite will librate around these equilibrium points. The ellipticity of the equator is represented by the degree two and order two sectoral coefficients ($C_{2,2}$ and $S_{2,2}$), which is further characterized by the equatorial moments and products of inertia of the Earth. Blitzer et al. (18) showed the existence of the equilibrium points and noted that geosynchronous satellites could be especially useful in determining $C_{2,2}$ and $S_{2,2}$. The ellipticity of the equator is a 100-meter undulation of the geoid at the equator. It is known now that the major axis of the elliptical equatorial cross-section is oriented at 15° west of the Greenwich meridian. From the dynamic point of view, the orientation of the principal axis determines where the Earth's equatorial products of inertia are zero.

Through the 1960s, numerous developments occurred in both theory and analysis. In 1966, Kaula published a landmark book on satellite geodesy. He developed a linear theory that described the influence on orbital motion that results from any term in the gravity field of the Earth. The theory enabled characterizing the perturbations in frequency space, that is, the frequency and period for changes in the satellite orbital elements produced by any degree and order set of coefficients could be readily determined. With this relationship, short-period effects could be removed from the orbital model to elucidate the long-term variations caused by, for example, odd-degree zonal harmonics. And even though the theory did not predict the existence of equilibrium solutions for a geosynchronous satellite, it identified the conditions for both shallow resonance (Kaula's linear theory is valid) and deep resonance (such as the geosynchronous problem).

As more satellites were launched in the 1960s, a revolution was also taking place in the accuracy of tracking systems. The U.S. Navy launched a system of satellites, initially known as the Navy Navigation Satellite System (NNSS) and later known as Transit, to support the navigation of its fleet of submarines. The tracking system was based on the apparent shift in satellite transmitter frequency associated with the relative motion between the satellite and a radio receiver. During the decade, a new application of lasers was also made, which produced an accurate distance measurement to satellites. Satellite laser ranging (SLR) is based on measuring the time required for a short laser pulse to travel from a ground-based transmitter to a satellite where it is reflected back to the

source. The initial experiments with SLR showed that the system could measure the distance from the station to the satellite as accurately as a few meters, but within two decades, the systems were measuring with better than centimeter accuracy.

In the 1960s, several satellites were launched that carried SLR reflector arrays (LRA, laser reflector array), that direct the photons from an illuminating laser back to the source. But the series of Explorer satellites known as GEOS were the first satellites after Vanguard I to focus on geodetic problems. These satellites carried a Doppler tracking system similar to the NNSS, but operating at different frequencies, and a LRA.

By the early 1970s, the gravity field of the Earth had been determined to degree and order 16 (19). The determination of gravity and the corresponding improvements in the shape of the Earth, as reflected by the geoid, were enabled by improvements in the applied technologies foretold by Project Vanguard. The undulations in the geoid introduced by the various degree and order coefficients were found at the few meter level, except for the terms noted previously.

Satellite Geodesy from 1970–2000. Still other geodetic satellites were launched in the 1970s whose primary objectives were the gravity field and shape of Earth. To minimize the influence of nongravitational forces (atmospheric drag and solar radiation pressure), these satellites were spheres of small cross-sectional area, but high mass, which gave them a low ratio of area to mass, an important parameter in the magnitude of a nongravitational force. The first such satellite, known as Starlette, was launched by the French space agency in 1975 into a 50° inclination orbit and an altitude of about 1000 km. The surface of the 24-cm diameter passive sphere was covered with laser corner cubes to support ground-based SLR tracking, and the core (constructed of uranium 238) gave the small satellite a mass of about 48 kg. In 1976, NASA launched a 60-cm diameter spherical satellite known as LAGEOS-I (LAsER GEODYNAMICS Satellite) into an orbit that had an altitude of about 5800 km, but whose area to mass ratio was comparable to that of Starlette. This satellite was also covered with laser corner cubes.

In the 1990s, both Starlette and LAGEOS-I were followed by launches of twin satellites into different orbit planes. The Starlette twin, known as Stella, was launched into an inclination of about 98° , and LAGEOS-II was launched into an inclination of 50° . In 1989, the Soviet Union launched spherical geodetic satellites, known as Etalon-I and Etalon-II, into high-altitude orbits similar to those used by the Soviet navigation satellite system, GLONASS. One additional spherical satellite that had a geodetic purpose was launched by Japan, it was known as Ajisai. Still another spherical geodetic satellite was placed into a low-altitude orbit from the Mir space station in April 1995. This German satellite from GeoForschungsZentrum Potsdam, known as GFZ-1, was destroyed on re-entry in 1999.

A new geodetic tool, a radar altimeter, was tested in Earth orbit on the space station precursor, known as Skylab, in 1974. For the first time, this altimeter directly measured the distance between a satellite-borne instrument and the surface of the Earth by emitting microwave radiation and measuring the time between transmission of the pulse and the arrival of the echo from the Earth's surface, also known as a "time of flight" measurement. The radar

altimeter carried on Skylab showed the technical feasibility of such instrumentation, and it hinted at the ability to measure directly the shape of the Earth in the ocean regions. Because space-borne radar altimeters are nadir pointed, it is apparent that the altimeter can be used to map the surface after removing the effects of orbital motion and other corrections, such as atmospheric delays.

The Geodynamics Experimental Ocean Satellite GEOS-3 was launched in 1975 to extend the Skylab experiment into a regular operation with radar altimetry. Tracking data from the GEOS-3 Doppler system and SLR tracking instrumentation, as well as the altimetry data, played a prominent role in determining parameters in updated versions of the gravity field, particularly the Goddard Earth Model series, known as GEM (20).

The GEOS-3 launch was followed by a sequence of satellites that carried radar altimeters: Seasat in 1978, Geosat in 1985, ERS-1 in 1991, TOPEX/POSEIDON in 1992, and ERS-2 in 1995. These altimeters were all designed for optimal performance over oceans, large inland seas, and lakes. The parameters that describe the oblate spheroid character of the Earth could be determined by a least-squares fit of an ellipsoid to the ever increasing volumes of altimeter data in the ocean areas that accounted for 70% of the Earth's surface. Whereas Newton's prediction about the oblate nature of the Earth was originally confirmed by arc measurements in Lapland and Ecuador, direct confirmation from most of the ocean areas was made by satellite altimetry. Beyond this aspect, satellite altimetry has revolutionized oceanography, as described by Fu and Cazenave (21). Images of the ocean surface created from altimeter data revealed bathymetric features, including seamounts and ocean trenches, gravitational reflections of the ocean bottom on the ocean surface.

By the start of the twenty-first century, based primarily on the body of geodetic data collected for 40 years, the International Association of Geodesy's ellipsoid parameters of the Earth were (22) $GM_E = 398600.4418 \text{ km}^3/\text{s}^2$, $J_2 = 0.0010826359$, $W_0 = 62636856.0 \text{ m}^2/\text{s}^2$, mean radius = 6378136.7 m, and $1/f = 298.25231$. Comparison of these parameters with earlier values, such as those presented by Kaula (17), could suggest that the Earth has changed its shape during the period from 1960 to 2000. There is no evidence to support this suggestion. What has taken place in this interval is a significant improvement in the observational accuracy, so that the changes in the Earth model parameters reflect a convergence toward more accurate values rather than a change in the physical Earth.

In the reflection of the shape of the Earth through the gravity field and the corresponding geoid, dramatic improvements have been achieved in both the accuracy and resolution of gravity (and geoid) models derived from altimeter data. The most common gravity fields in use in 2001 were JGM-3 (23), EGM-96 (24), and GRIM-4 (25). As one representative model of the shape of the Earth, Fig. 1 shows the geoid computed from the gravity model of JGM-3, based on coefficients to degree and order 70, which corresponds to a horizontal resolution of about 600 km. This figure shows a "dip" in the geoid in the Indian Ocean of about 115 meters with respect to the reference ellipsoid. Likewise, a rise in the surface can be seen in the North Atlantic region. The lack of strong correlation between the geoid and surface topography is evident, although some correlation exists, such as in the Andes, because the satellite data have been augmented by

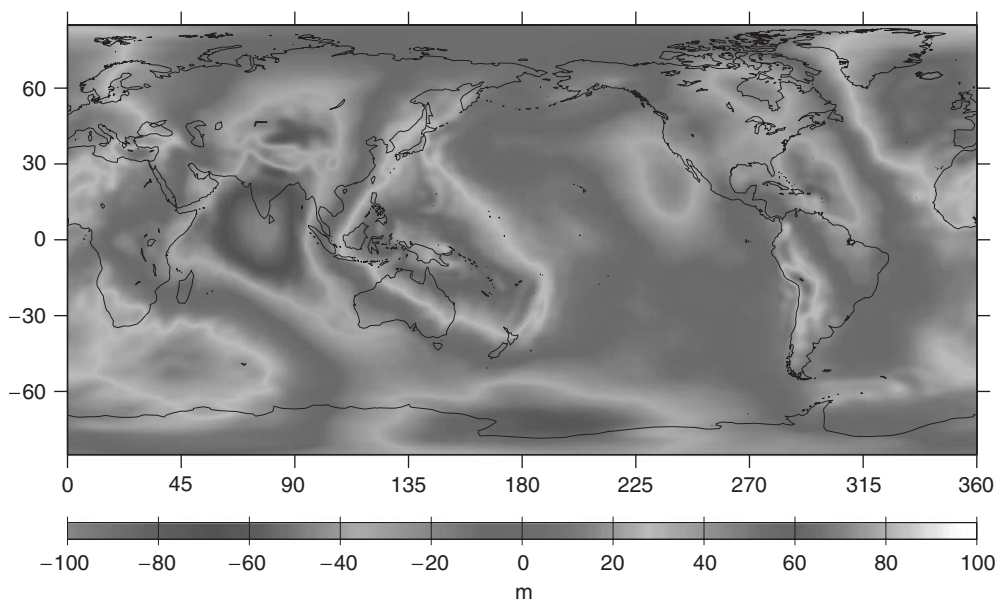


Figure 1. The shape of the Earth is illustrated by the geoid, a surface determined from the JGM-3 gravity field of the Earth. The geoid, which is a reflection of the mass distribution within the Earth, coincides with mean sea level in the ocean regions. The figure shows departures of the geoid from an oblate spheroid, or ellipse of revolution. Though the geoid is a suitable surrogate for the physical shape of the Earth in the ocean areas, local land topography (such as Mt. Everest) must be added to represent fully the topographic characteristics in landmass areas. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

surface gravity measurements. The difference between the geoid and surface topography in the Indian Ocean, for example, can be explained by the reminder that the geoid represents the shape of the gravity field, which, in turn, depends on subsurface mass distribution that does not always correlate well with surface features on the Earth.

In summary, the shape of the Earth cannot be described with high accuracy using a simple geometric figure. At least in the ocean areas, the mean sea surface closely coincides with the geoid (within ± 125 meters everywhere). The instantaneous ocean surface departs from the geoid because of temporal effects (e.g., tides and meteorology) at the few meter level. For landmasses, the physical topography of the surface exhibits large variations in elevation from sea level to Mt. Everest (9 km) and other striking changes in elevation over short horizontal distances (e.g., the Grand Canyon). Nevertheless, the geoid surface is contained within ± 125 meters of the oblate spheroid given by Groten (22), even in land areas. For applications that require a highly accurate representation of the surface topography instead of the geoid surface, accounting for the local land topography can be applied, such as a 1-kilometer resolution Digital Elevation Model (26).

The Future. The confluence of several technologies in the late 1950s enabled dramatic improvements in describing the Earth's shape: artificial satellites,

networks of ground-based satellite tracking systems, space-borne radar altimeters, and high-speed computers. By the end of the twentieth century, new technologies had been developed to support the ever increasing accuracies required for scientific investigations.

Applications of satellites to the determination of the size and shape of the Earth were enabled by the Minitrack system at the beginning of the space age, but by the beginning of the twenty first century the Global Positioning System (GPS) was playing an increasingly prominent role. GPS receivers carried on satellites in low Earth orbit (LEO) provide global and continuous observations of the orbital perturbations introduced by the Earth's gravity field and other forces. The GPS receiver carried on TOPEX/POSEIDON contributed to the JGM-3 gravity model, for example, and such receivers carried on other LEO satellites will contribute to further improvements in knowledge of the size and shape of the Earth. Geodesy itself is undergoing a revolution because of GPS, with a variety of applications in the Earth sciences that require determination of position with an accuracy at the centimeter-level or better.

The most detailed descriptions of the topography of Mars were obtained by using a laser altimeter carried on the Mars Surveyor in the closing years of the twentieth century. The altimeter, known as MOLA (Mars Orbiter Laser Altimeter), provided the most detailed topographic map of Mars to date. Although similar in principle to the "time of flight" radar altimeter measurement, the narrow beam laser allows surface averaging over much smaller areas within the beam, an important factor over rough terrain. Analysis of MOLA data showed a dramatic difference in surface elevation between the high Southern and low Northern Hemispheres of Mars (27), which suggests a dichotomy in the evolution of each. In terms of a simple geometric figure, Mars can be reasonably described as a triaxial ellipsoid that has 2- to 5-kilometer differences in the axes. The center of figure is offset from the center of mass by 3 km in the polar axis direction. The solar system is ripe for applications to other celestial bodies in the twenty-first century.

Mapping of the Earth's topography is undergoing a revolution with instrumentation that provides improved horizontal and vertical resolution. In 2000, a Space Shuttle mission carried the Shuttle Radar Topography Mission, a radar interferometric system designed to map most of the land surface between 56°S and 60°N. A digital elevation model that has 30-m horizontal resolution and 15-m vertical accuracy will be produced from the data (28).

Whereas the geodetic emphasis in the first two decades of the space age was on characterizing the Earth's shape, the emphasis at the end of the twentieth century was moving toward detecting topographic *change* and scientific interpretation of the changes. The twenty-first century will bear witness to the application of new technologies to detect change at even greater accuracy. For example, the Geoscience Laser Altimeter System will begin operations in 2003 on the Ice, Cloud and Land Elevation Satellite (ICESat) to provide highly accurate elevations of the land surface, especially to detect topographic change in the major ice sheets of Greenland and Antarctica (29). The growth or diminishment of the ice sheets has a direct relation to sea level and climate change.

It took almost 2000 years to refine human knowledge of the shape of the Earth from a sphere to an ellipsoid of revolution. After the first artificial satellites

were launched, the next refinement represented by the pear shape took only two years. In successive years since the first artificial satellites were launched, the shape of the Earth has been refined into high-resolution digital models. Determinations of temporal changes in the topographic characteristics that were begun in the late twentieth century will blossom in the twenty-first century as new and more accurate measuring tools are developed. The tools will provide the fundamental measurements, but scientific interpretation of temporal changes in topography and the relation of topography to other phenomena, such as the evolution of El Niño in the Earth's oceans and the surface evolution of celestial bodies, will be the areas emphasized in the twenty-first century.

BIBLIOGRAPHY

1. Singer, F. *A Short History of Scientific Ideas to 1900*. Oxford University Press, Oxford, England, 1959.
2. Woolard, E. *Astronomical Papers*, Volume XV, Part I. U.S. Naval Observatory, Washington, DC, 1953.
3. Newton, I. *Philosophiae Naturalis Principia Mathematica*. London, 1687 (Dated July 1686; translated into English by A. Motte, 1729, with revisions by F. Cajori, University of California Press, 1934).
4. Todhunter, I. *A History of the Mathematical Theories of Attraction and the Figure of the Earth*, Macmillan, London, 1873 (republished by Dover, New York, 1962).
5. Butterfield, A. *A History of the Determination of the Figure of the Earth from Arc Measurements*. The Davis Press, Worcester, MA, 1906.
6. Smith, J.R. *Introduction to Geodesy*. Wiley, New York, 1997.
7. Jeffreys, H. *The Earth*, 3rd ed., Cambridge University Press, Cambridge, England, 1952.
8. Kaula, W. *Satellite Geodesy*, Blaisdell, Waltham, MA, 1966 (republished by Dover, New York, 2000).
9. King-Hele, D. The Earth's gravitational potential, deduced from the orbits of artificial satellites. *Geophys. J. R. Soc.* V. 4: 3–16 (1961).
10. Szebehely, V., and H. Mark. *Adventures in Celestial Mechanics*. Wiley, New York, 1998.
11. Merson, R., and D. King-Hele. Use of artificial satellites to explore the Earth's gravitational field: Results from Sputnik-II. *Nature* 182: 640–641 (1958).
12. Green, C.M., and M. Lomask. *Vanguard – A History*. NASA SP-4202, Washington, DC, 1970.
13. Stehling, K., *Project Vanguard*. Doubleday & Company, New York, 1961.
14. O'Keefe, J., A. Eckels, and R. Squires. Vanguard measurements give pear-shaped component of Earth's figure. *Science* 129: 565–566 (1959).
15. O'Keefe, J., A. Eckels, and R. Squires. The Gravitational field of the Earth. *Astron. J.* 64 (7): 245–253 (1959).
16. Veis, G (ed.). *The Use of Artificial Satellites for Geodesy*. Wiley, New York, 1963.
17. Kaula, W. A geoid and world geodetic system based on a combination of gravimetric, astrogeodetic, and satellite data. *J. Geophys. Res.* 66 (6): 1799–1811 (1961).
18. Blitzer, L., E. Boughton, G. Kang, and R. Page. Effect of ellipticity of the equator on 24-hour nearly circular satellite orbits. *J. Geophys. Res.* 67: 329–335 (1962).
19. Gaposchkin, E.M., and K. Lambeck. Earth's gravity field to the sixteenth degree and station coordinates from satellite and terrestrial data. *J. Geophys. Res.* 76: 4855–4883 (1971).

20. Lerch, F., S. Klosko, R. Laubscher, and C. Wagner, Gravity model improvement using Geos 3 (GEM 9 and 10). *J. Geophys. Res.* 84: 3897–3916 (1979).
21. Fu, L., and A. Cazenave (eds). *Satellite Altimetry and Earth Sciences*. Academic Press, San Diego, 2001.
22. Groten, E. Report of the International Association of Geodesy Special Commission SC3, Fundamental Constants, XXII Gen. Assembly Int. Union Geodesy Geophys. Birmingham, (U.K.), 1999.
23. Tapley, B., M. Watkins, J. Ries, G. Davis, R. Eanes, S. Poole, H. Rim, B. Schutz, C. Shum, R. Nerem, F. Lerch, J. Marshall, S. Klosko, N. Pavlis, and R. Williamson. The JGM-3 geopotential model. *J. Geophys. Res.* 101 (B12): 28029–28049, 1996.
24. Lemoine, F., S. Kenyon, J. Factor, R. Trimmer, N. Pavlis, D. Chinn, C. Cox, S. Klosko, S. Luthcke, M. Torrence, Y. Wang, R. Williamson, E. Pavlis, R. Rapp, and T. Olson. The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96. NASA/TP-1998-206861, Washington, DC, 1998.
25. Schwintzer, P., C. Reigber, A. Bode, Z. Kang, S. Zhu, F. Massmann, J. Raimondo, R. Biancale, G. Balmino, J. Lemoine, B. Moynot, J. Marty, F. Barlier, and Y. Boudon. Long-wavelength global gravity field models: GRIM4-S4, GRIM4-C4. *J. Geodesy* 71: 189–208 (1997).
26. Hastings, D., and P.K. Dunbar, Global Land One-kilometer Base Elevation. NGDC Key to Geophysical Records Documentation No. 34, NOAA, Boulder, CO, May 1999.
27. Smith, D.E., M. Zuber, S. Solomon, R. Phillips, J. Head, J. Garvin, W. Banerdt, D. Muhleman, G. Pettengill, G. Neumann, F. Lemoine, J. Abshire, O. Aharonson, C. Brown, S. Hauck, A. Ivanov, P. McGovern, J. Zwally, and T. Duxbury. The global topography of Mars and implications for surface evolution. *Science* 284: 1495–1503 (1999).
28. Farr, T., and M. Kobrick. Shuttle radar topography mission produces a wealth of Data. *Eos Trans. Am. Geophys. Union* 18 (28): 583 and 585 (2000).
29. Zwally, J., B. Schutz, W. Abdalati, J. Abshire, C. Bentley, A. Brenner, J. Bufton, J. Dezio, D. Hancock, D. Harding, T. Herring, B. Minster, K. Quinn, S. Palm, J. Spin-hirne, and R. Thomas. ICESat's laser measurements of polar ice, atmosphere, ocean and land. *J. Geodynamics*, Special Issue on Laser Altimetry, in press.

B.E. SCHUTZ
 Center for Space Research and
 Department of Aerospace Engineering
 and Engineering Mechanics
 University of Texas at Austin
 Austin, Texas

SKYLAB

Introduction

A space station has been an obvious part of space exploration since people began to think seriously about space travel. In 1923, Hermann Oberth published *The Rocket into Interplanetary Space* (1), which contained the first serious proposal for a manned space station to appear in scientific literature rather than fiction. Oberth suggested a permanent station supplied by smaller rockets, and he

suggested rotation of the vehicle to produce artificial gravity for the crew. "Such a station," he said, "could serve as a base for Earth observations, as a weather forecasting satellite, as a communications satellite, and as a refueling station for extraterrestrial vehicles launched from orbit." He wrote from an engineer's viewpoint, leaving out scientific research, but he included just about every other use for a station proposed since then.

The *First Symposium on Space Flight* was held on 12 October, 1951 at the Hayden Planetarium in New York City. Papers read at the symposium were published by Collier's Magazine beginning in March 1952 and continuing to April 1954 under the title, "Man Will Conquer Space Soon." Contributors included Wernher von Braun, who laid out a manned space exploration program starting with sorties, progressing rapidly to a space station, and reaching the Moon in the year 2000 (2). Then, in 1959, von Braun advanced a theory for using a booster's spent stage as a space station's basic structure. This later evolved into the "wet workshop" concept for what became Skylab, the world's first space station.

The Cold War turned von Braun's logical progression on its head. In 1959, the newly formed NASA was already propounding a lunar landing as a long-term goal, and in 1962, President Kennedy proclaimed it as the goal for NASA. A space station would have to wait.

But the goal remained, and as spacecraft and boosters for Project Apollo, NASA's lunar landing program, were defined, space station concepts derived their specifics from them. Modifications of the Command Module, the Lunar Module, and various configurations of the Apollo booster series, were studied. NASA wanted to use its hardware for programs beyond Apollo. In 1965, the concept of Apollo Extension Systems was formalized at NASA Headquarters and very soon became the Apollo Applications Program (AAP). The concept later named Skylab was just one of the missions proposed by AAP (3).

Development

Skylab was a program that was being designed and trained for simultaneously, and the design and training were taking place in parallel with the Gemini, and later, the Apollo manned flights. Thus, the problems and experiences of those flights shaped the Skylab design.

Early NASA flights were conducted amid a sense of mystery as to how and whether humans would withstand the enigmatic environment of weightlessness and the confines of a tiny spacecraft. As the results came in, the astronauts seemed to hold up pretty well. But weight loss, loss of blood volume and muscle strength, some vestibular disturbances, and the suspicion that bones were thinning left unknown our ability to remain healthy and productive on long flights.

Answering this question was to be Skylab's first goal. The second goal was to define the engineering arrangements needed for an effective human hotel/laboratory in space. The third was to accomplish and demonstrate real first-class scientific returns in three fields: the study of the Sun, of Earth, and of the behavior of materials in weightlessness itself.

In 1966, these goals were still vague. AAP was still a rather primitive concept of using Apollo's Saturn booster as an experimental space station. As Saturn evolved, one of its upper stages featured a restartable liquid hydrogen/liquid oxygen engine. Called the "S-IVB" (pronounced "Ess Four Bee"), it was the third stage of the Saturn V launch vehicle; it burned briefly to achieve Earth orbit, then restarted to boost the Command Service Module (CSM) and Lunar Module (LM) out of Earth orbit toward the Moon. It also formed the second stage of the smaller S-IB booster used in the earliest Apollo missions. It appeared to be a good candidate for manning (Fig. 1).

Between 1961 and 1965, the Marshall Space Flight Center (MSFC) conducted several studies to define the design and uses of space stations. In 1965, a formal study was begun of the "S-IVB Spent Stage Workshop." "Spent stage" meant that the S-IVB was to be used as a working booster, and after its liquid hydrogen fuel was spent, the fuel tank would be vented, pressurized with oxygen, and used by the crew. The crew would dock with the spent stage, open the

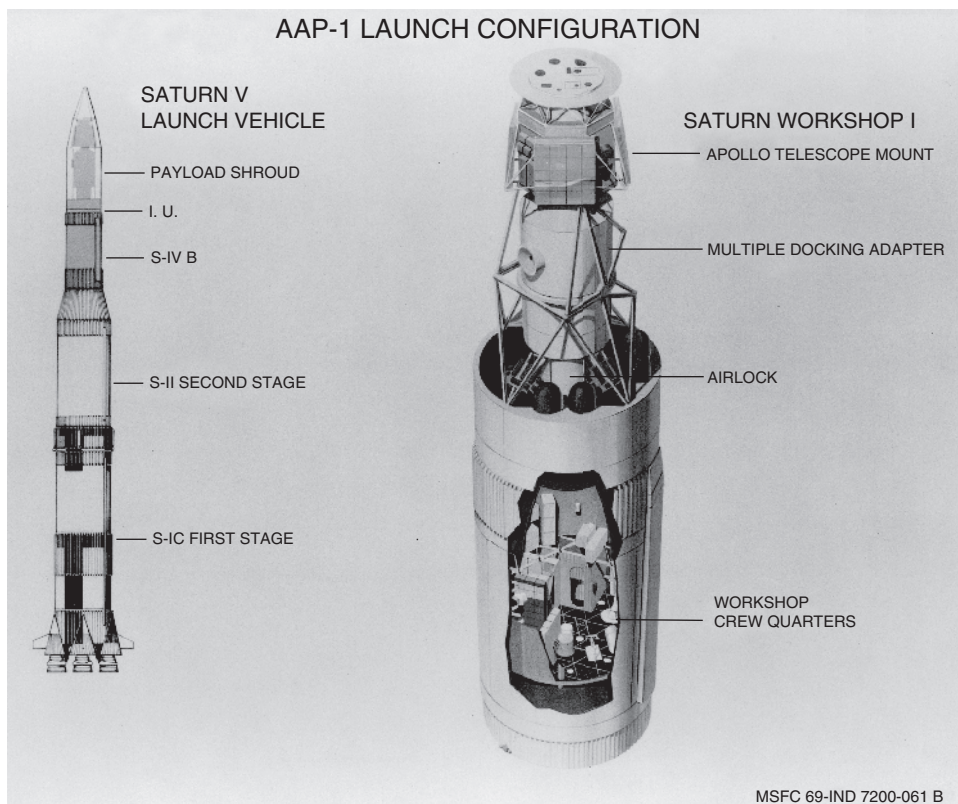


Figure 1. The Saturn V rocket launch configuration used to launch the Skylab module on the first Apollo Applications Program Flight (AAP-1) and also a cutaway drawing of the modified S-IVB that became Skylab. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

48 large bolts that held the hatch in place, enter the stage, and spend time living in and evaluating the roomy but sparsely furnished volume.

On 19 August, 1966, George E. Mueller, NASA Associate Administrator for Manned Space Flight, used a felt pen and poster paper to pin down the final conceptual layout for the budding space station's major elements (Fig. 2). In 1967, a four-flight AAP program was outlined that included two manned flights to the "Spent Stage Workshop."

Of course, the engineers at Marshall Space Flight Center wanted to outfit the workshop as fully as possible before launch. But the challenge of providing wiring and other equipment that could withstand a bath of several hours' duration in liquid hydrogen and then function reliably was proving difficult and limiting.

Suddenly, that changed, and a decision that was bad for Apollo turned out to be good for Skylab. The last three manned visits to the Moon, Apollos 18, 19, and 20, were canceled for budgetary reasons. Their boosters, the mighty three-stage

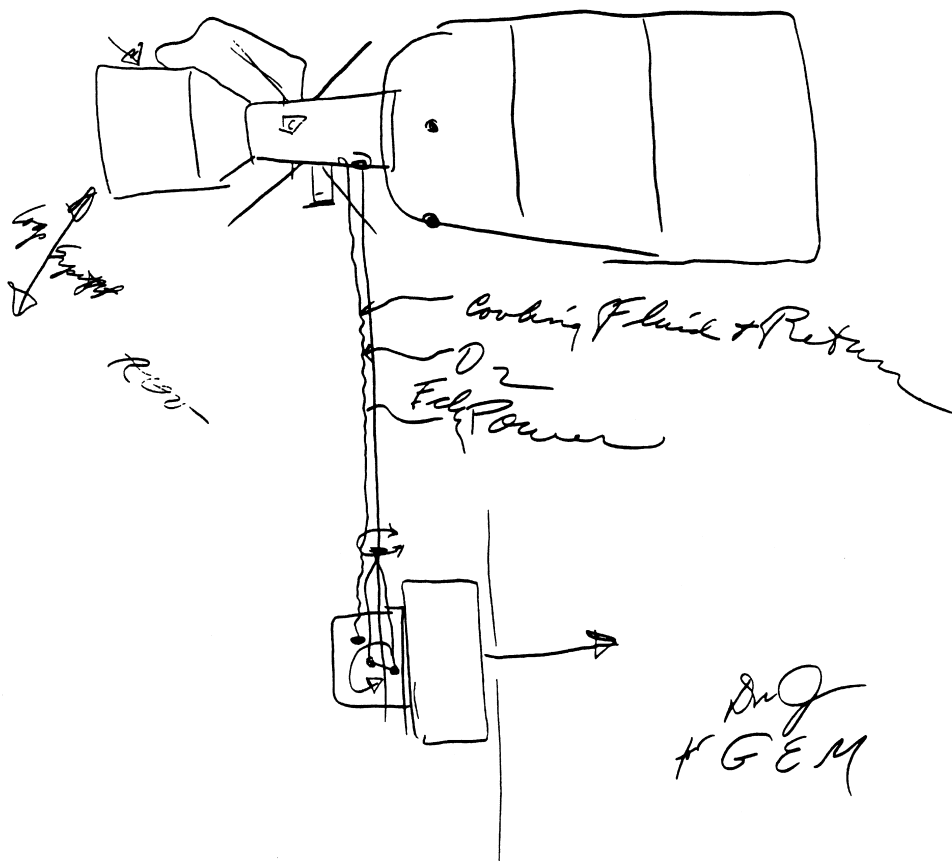


Figure 2. The sketch made by Dr. George Mueller, NASA Associate Administrator for Manned Space Flight, on 19 August, 1966 to illustrate the basic components of what eventually became Skylab.

Saturn Vs, became available. A Saturn V used its third stage, the same friendly S-IVB, to launch the Apollo spacecraft from Earth orbit to the Moon. Just to go to Earth orbit did not require it to function at all as a booster. It could be launched dry, and all the needed equipment would be preinstalled. There was lots of weight available. The Dry Workshop was born (Fig. 1).

The new configuration allowed a significant upgrade of the medical experiment complex. Real in-flight data could now be collected. The major experiments became an exercise tolerance experiment using a bicycle ergometer and a mass spectrometer to measure oxygen consumption and carbon dioxide production; lower body negative pressure to measure the crew's probable response to reentry at end-of-mission; vestibular experiments featuring a rotating chair to stimulate and characterize space motion sickness; an electroencephalogram (EEG) to measure the duration and effectiveness of sleep; and a complete and accurate intake and output study that involved measuring and sampling every drop of urine, drying and returning feces, and accounting for every gram and calorie of food. Blood would be drawn weekly.

Another AAP mission being developed used the Apollo Lunar Module (LM) as the core for a solar physics mission. A cluster of cameras and telescopes optimized to study the Sun above the blockage and distortion of Earth's atmosphere would be mounted in the LM descent stage; the controls and displays would be in the ascent stage where Apollo crews normally lived. The crew would arrive in an Apollo Command/Service Module (CSM), dock, perform experiments, and then return to Earth in the same CSM. The modified LM, called the "Apollo Telescope Mount," ATM for short, generated a lot of scientific enthusiasm. But as the 1960s progressed, it became apparent that there was not going to be enough money to carry it out. When the Dry Workshop and its Saturn V booster and huge weight capacity became available, the ATM (without a separate ascent stage) was grafted onto it, and it became the scientific crown jewel of Skylab.

The Earth Resources Experiment package was a relative latecomer, but a good one. The idea was to use the crew's eyes and hands to select and point high-resolution, multispectral cameras at all kinds of phenomena on land and ocean—meteorological, crop, habitation and pollution, and natural phenomena such as volcanic eruptions. A capable set of cameras was located on the Earth-pointing side of the docking adapter (a cylindrical section that connected the Workshop with the docked CSM and also contained the ATM controls and displays and a lot of systems equipment).

The final major addition was the materials science package, notably a furnace for melting metal, creating alloys, and studying these processes in the absence of gravity-induced convection. In addition, literally dozens of smaller individual experiments were added to the mix. Some were devised by students at American schools.

First Mission

After many delays, the launch of the Skylab Orbital Workshop (OWS) was scheduled for 14 May, 1973, to be followed 24 hours later by the launch of the first crew (Fig. 3). The crew was commanded by Navy Captain Charles Conrad, Jr.,

a very experienced astronaut, who had flown a previous record-setting long duration flight (Gemini V for 8 days), then commanded Gemini X and Apollo 12, the second lunar landing flight. (A rather short man, he was famous for his stirring words on that occasion: “This may have been a small step for Neil, but it’s a giant leap for me!”) His crewmates were this author, Joe Kerwin, a Navy flight surgeon, and Paul Weitz, a Navy test pilot.

The crew watched the Workshop launch from the roof of their training building, apparently a complete success. But at about 36 seconds into the launch, mission control observed a g-spike, a brief but heavy jolt to the booster just as it

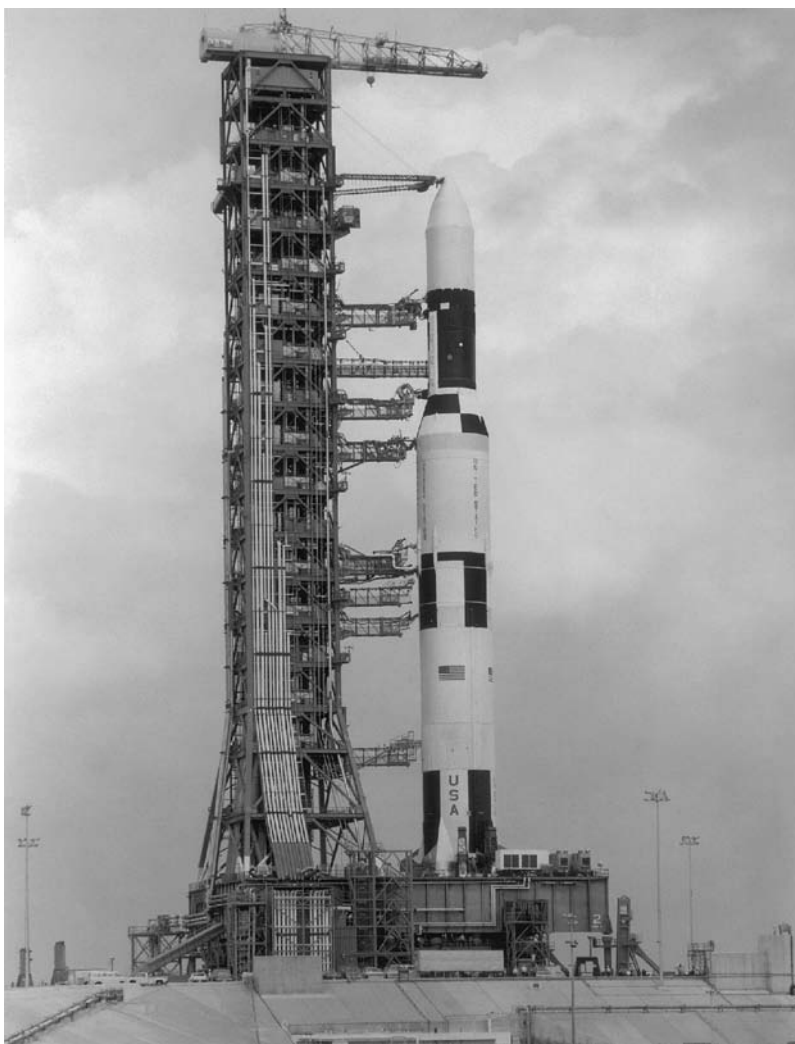


Figure 3. The Skylab Module ready for launch on 14 May, 1973. Note the absence of the escape tower at the top of the stack. Because no people were on board, the escape rocket was deemed unnecessary. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

went supersonic. Launch continued without any apparent problems; the correct circular orbit of 235 nautical miles was achieved.

The problems were revealed when the deployment sequence of the Workshop was attempted. First, the "Apollo Telescope Mount," the complex of solar physics experiments that had its own set of solar panels for electrical energy, was commanded to rotate out 90° and deploy its solar panels. This operation was completed. Next, the two large solar panels on the Workshop needed to be deployed. They were accordion-folded inside a pair of covers, which in turn were hinged and folded down flush with the sides of the OWS. Beneath them was the last item needing deployment: a cylindrical heat/micrometeoroid shield made of stout aluminum, wrapped around the OWS side walls. Springs were to expand and deploy it about 6 inches away from the OWS walls, where it would reflect solar energy away and break up micrometeorites.

When the OWS solar panels were commanded to deploy, there was no response from the starboard panel at all. A tiny trickle of current was seen from the port panel, indicating partial deployment. Attempts to deploy the heat shield also brought no response. And when temperatures both outside and within the OWS began to rise sharply, the problem became obvious; the heat shield had deployed prematurely, during launch, and was gone.

The crew's launch the next day was scrubbed, and the Skylab team evaluated the situation. There were two problems:

1. The absence of the heat shield was causing intolerable temperatures within the OWS, up to 135°F.

2. One of the two main solar panels appeared to be gone, and the other was stuck (mostly) closed. Two thirds of Skylab's power was unavailable; only the solar panels on the ATM remained. Skylab's systems and experiments would be crippled.

The engineering team immediately began developing jury-rigged solutions to the loss of the heat shield; in 10 days, three solutions were designed, built, tested, and loaded aboard the first crew's spacecraft. Another team evaluated possible causes for the stuck solar panel and selected tools for the crew to use in an attempt to free it. Meanwhile, the flight control team worked to keep the workshop's temperature under control by rotating it away from the Sun—but not so far as to lose all power from the remaining solar panels. Other teams assessed the systems status and medical safety of the damaged OWS and rewrote the mission rules.

Skylab, in an unusually high-inclination orbit (50°, allowing Earth photography across most of the populated land mass), passed directly over Cape Kennedy every 5 days. So, the first new opportunity to launch the crew into an orbit suitable for rendezvous was 20 May, 1973. But the repair hardware was not ready. Launch was set for 25 May, and it was successful.

Rendezvous with Skylab took place 8 hours after launch, and the crew visually confirmed the suspicions of the team about damage. Only one heavy aluminum strap, a fragment of the missing heat shield, seemed to be holding down the remaining solar panel. So the crew did an immediate "space walk" from the command module side hatch; Weitz, with a shepherd's crook, tried to pry the panel loose. But the strap was too strong. So, they contented themselves with

taking and transmitting to Houston extensive television pictures of the damage and then (after a second EVA to fix the docking system) docked with Skylab.

On day 2, the first substitute heat shield was deployed through an experiment hatch on the sunny side of the Workshop—an ingenious “parasol,” built from telescoping fishing poles and lightweight nylon cloth. It worked; temperatures in the OWS began to drop. By the third day, it was cool enough to begin activation of workshop systems (Fig. 4).

The first 2 weeks of the first mission were conducted in semidarkness, as the crew performed what experiments they could while conserving precious power. Most lights were kept off; no hot drinks were allowed. The Earth resources experiments, which required the Skylab cluster to deviate from its normal attitude facing the Sun to one in which it tracked Earth’s surface, were abandoned after one disastrous attempt caused depleted batteries to drop off line. The medical work and some solar physics could be accomplished.

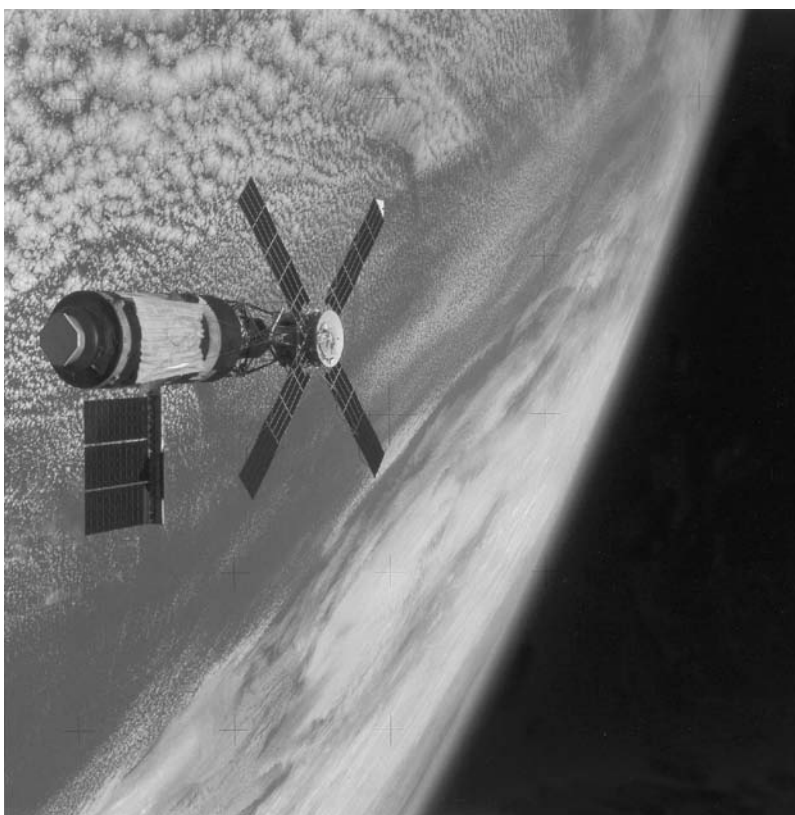


Figure 4. The final configuration of Skylab after the repair of the remaining solar panel and its deployment. The solar panel that should have been on the opposite side of the Spacelab vehicle was accidentally torn off when the shroud that covered the payload was lost during the launch sequence. The solar shield that prevented the spacecraft from getting too hot is the light colored area covering Skylab. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The Johnson Space Center engineers in Houston were working on a plan, and by day 10, they were ready to describe to the crew a recommended method for deploying the stuck solar panel. As anticipated, it required a space walk. No external maintenance had been planned for Skylab; the space suits were there only to retrieve and replace film from the solar telescope/cameras. But they were there, and the needed tools had launched with the crew: 25 feet of aluminum poles in 5-foot sections; a telephone company tree-limb lopper with lots of rope; and extra hooks and tethers. Even a dental saw from Skylab's medical kit was taped to one suit, just in case. There were no handrails or footholds in the damaged area and no lighting.

Conrad and Kerwin went out on day 14, freed up the solar panel cover, and watched in joy as the accordion-packed panels, now exposed to the warming Sun, expanded. The power shortage was ended.

The remaining 2 weeks of the mission were a little too short to catch up completely on all of the planned experiments. But with power everything worked, and it was obvious that there would be a second and a third Skylab mission.

Splashdown in the Pacific Ocean west of San Diego occurred on 25 June within a mile of the waiting U.S.S. Kearsarge. For the first time, the intact command module was hoisted onto the carrier's flight deck before the crew egressed, reducing the risk that a crew member weakened by long exposure to space would be unable to egress on the ocean and use the helicopter's hoist. The crew was weary, but in good condition and spirits, able 3 days later to meet President Nixon and Soviet Premier Brezhnev in San Clemente.

Medically, the first crew

1. had lost from 5–10% percent of body weight;
2. had lost about 30% of arm strength and nearly 60% of leg strength; and
3. had lost about 10% of red cells from their blood, with no sign of new production.
4. by gamma ray imaging, had not lost appreciable bone calcium even after a month in space (although the very accurate intake–output study showed a net loss of calcium from the body).

Most importantly, the in-flight medical experiments showed a leveling-off effect after the first 2 weeks. Ability to exercise on the bicycle ergometer was fully maintained (muscles that were used did not weaken), weight loss did not continue, and performance and well-being were maintained. There was no reason not to continue the plan and expose the second crew twice as long.

Second Mission

On 28 July, 1973, the second crew launched to Skylab. Navy Captain Alan Bean, the commander, was a veteran of the Apollo 12 Moon landing. This was his first mission in command. His crewmates were Owen Garriott, a Ph.D. electrical engineer, and Jack Lousma, a Marine Corps test pilot.

Unfortunately, space motion sickness, which by the luck of the draw had been absent on the first mission, struck two of the three members of the Skylab II crew. Also, it remained true that activating a large space station was still a complex activity and required significant learning to handle oneself and one's equipment, no matter how well trained one was. The result was that by the end of the first week Bean, Lousma, and Garriott found themselves healthy but behind their time line.

They pledged to recover by working overtime, if necessary, and they certainly did. The time required to perform experiments and maintenance activities decreases with repetition, and they worked at it long and hard. Dr. Garriott, the science pilot, declined to use the experimental shower in favor of additional runs of solar physics and student experiments.

The same disciplined approach was applied to crew exercise—increased over the first crew's routine—to Earth photography and solar physics data collection, all of which were accomplished without incident. Data results from Skylab II were greater than planned, and in Houston, the Mission Control Team congratulated itself that it had finally learned how to schedule Skylab for optimum performance.

The accomplishments of this mission also included greatly expanding the scope of external repairs and maintenance activity conducted during three space walks and helping Arabella, the spider, to construct a web (she succeeded on the second try.)

From a medical perspective, the second crew returned from 59 days in space in as good condition as the first crew had after half as long. The increased exercise had not reversed the loss of muscle strength but had prevented it from going any deeper. Similarly, the loss of red cell mass actually recovered a bit after 4 weeks of decrease; their bodies were manufacturing erythrocytes again. Post-flight measurements showed that bone density had decreased linearly; that was still a long-term problem. Crew morale and performance and crew-ground teamwork were impeccable. NASA was justified in making the real-time decision to go for 84 days for the final Skylab crew.

Third Mission

Marine Corps Colonel Gerry Carr, scientist Ed Gibson, and test pilot Bill Pogue launched to the Skylab complex on 16 November, 1973, with plans for a third and final mission up to 84 days long—enough to “certify” human stays of 3 months on a future space station or interplanetary trip.

Certain things in the background are important to understand the challenge and events of Skylab III:

1. Vietnam and the social revolution of the Baby Boomers had taken center stage in America; the supreme excitement of space exploration and its Cold War backdrop had abated. What was the message for Skylab? It was that there was not going to be a “Skylab B” or any kind of near-term follow-up mission. The backup Skylab workshop would launch to the Smithsonian (where it can still be found), and scientific investigators knew that if any

more data were to be wrung from this magnificent assortment of instruments, the third mission was their last chance.

2. NASA's sole remaining new manned program was the Space Shuttle. NASA had argued that future space exploration of any kind necessitated lower cost access to Earth orbit; that this required a reusable vehicle; and that the fastest, most reliable way to design such a vehicle was to make it piloted. The implication was that we had shown that people could perform skilled tasks well in space and that DSMS (dreaded space motion sickness) was not going to stop them.

The third crew knew that the second crew had experienced motion sickness, and they had picked up the strong impression that the Shuttle program would prefer not to have a repetition. So, when one of them became motion sick and vomited during the first day, they discussed the matter on the intercom and decided not to bother Mission Control with it.

Unfortunately, they forgot that the intercom was routinely tape recorded and dumped to the ground. The result was a more or less public rebuke to the crew by their boss, Alan Shepard, which was not a morale-building way to get started on human's longest voyage in space.

Add to this the fact, noted before, that Mission Control had learned how to schedule crews on Skylab for densely packed, productive workdays. What they did was to forget the steep learning curve both previous crews had demonstrated, and they began to overschedule Jerry Carr's crew. The harder they worked, the "behinder" they got; the motion sickness incident hampered communications; and it was a few weeks into the flight before good communication took place and everybody got on the same page.

The rest is history. A repeat of history, actually; Skylab III followed Skylab II's path to high productivity. The crew exceeded all experiment goals, photographed the comet Kohoutek, observed the transit of Mercury across the Sun with instruments from the ATM, celebrated Christmas 250 miles up, grew beards, and in all things did honor to themselves and the Skylab team. They had been given more control over their time line and ran a "shopping list" of additional experiments under their own control. This is the way to run a space station.

How did they do medically? Very well, indeed. They exercised even more than Bean's crew. Because it was noticed that leg strength was still being lost in space despite bicycle ergometer and isometric exercise, a treadmill was deemed needed. A conventional treadmill could not possibly be carried up in the space-limited CSM. It fell to Dr. William Thornton, a physician-engineer-astronaut who would not fly until the Shuttle was operational, to design one made of a sheet of slippery Teflon. The Teflon was pinned to the floor of the workshop near the bicycle. The rejected bicycle harness was found and taken out of stowage. The crew member used it to strap himself to the floor over the teflon strip, his body pitched forward about 30°, holding a handhold on the wall. Then wearing cotton socks, he ran in place (Fig. 5). It was not ideal, but it worked. The third crew came back from space with stronger muscles and less weight loss (Carr even gained a pound!) than any of their predecessors. Long duration human space-flight had been proved both practical and productive (Figs. 6 and 7).

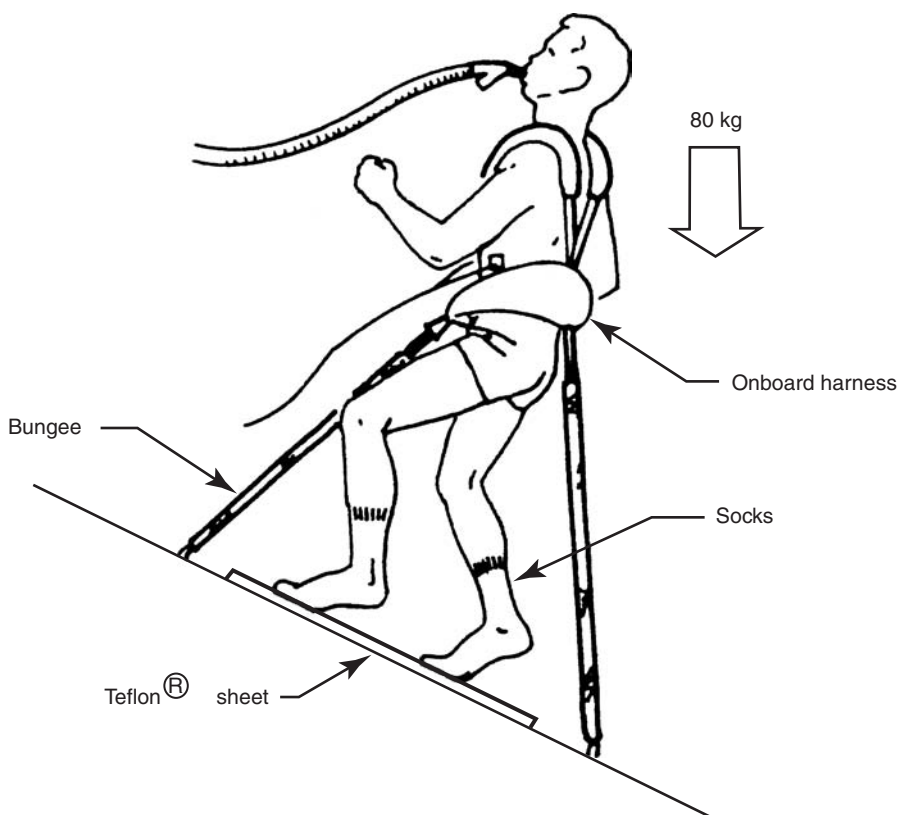


Figure 5. This drawing shows the Skylab treadmill devised by Astronaut Dr. William Thornton, a physician who is also an engineer.

It is worth noting that no significant illnesses or injuries occurred on any of the Skylab missions. The crew was, of course, rigorously screened for good health, so the odds were against a heart attack, stroke, or other major medical catastrophe. But the fact remains that whenever enough people spend enough time in space, there are going to be serious illnesses to deal with. Learning to do this is one of the goals of the International Space Station.

Summary

Skylab's goals were to discover whether humans could live safely in weightlessness for 3 months, to uncover and solve any design problems living in space might present, and to demonstrate that productive science experiments could be accomplished. Here is a look at the results.

Medical. We have reviewed the results from each mission. We will now summarize the medical view of human spaceflight as seen at the conclusion of Skylab in 1974:

- Middle-aged male humans do quite well in space for as long as 3 months (and there is no reason to suppose that females will not do just as well).

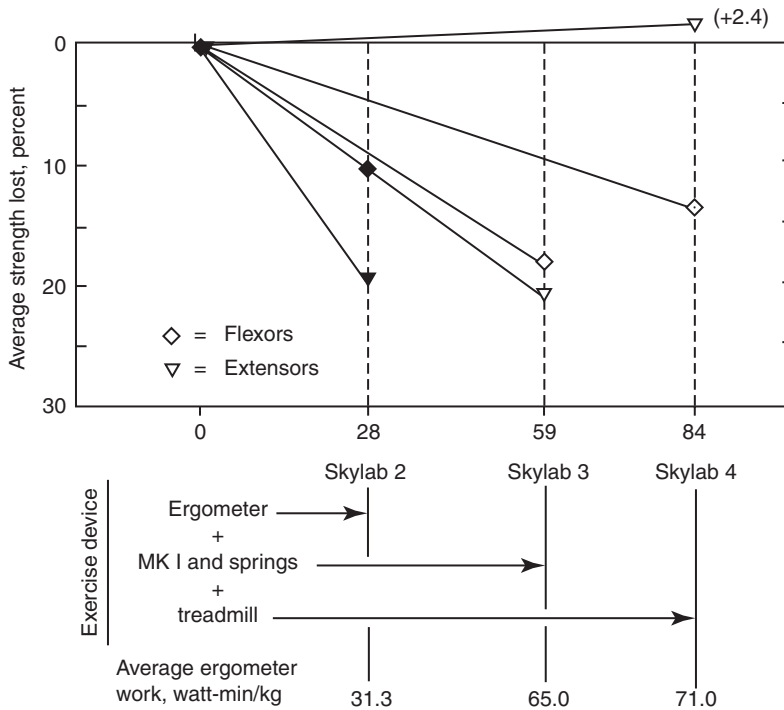


Figure 6. The graph illustrates the changes in average leg strength experienced by the Skylab astronauts during the three missions. It clearly shows the loss of leg strength in short missions from minimal exercise and then the maintenance of leg strength on longer missions from more intense exercise. The lower panel is a comparison among the various exercise devices that were used.

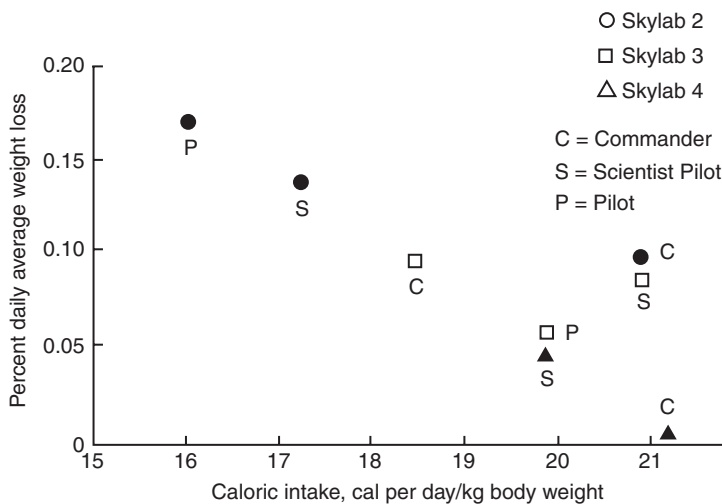


Figure 7. This graph shows the weight changes experienced by the nine Skylab astronauts as a function of caloric intake.

- However, weightlessness is stressful. It causes changes in our bodies, and although all are “adaptations” to the lack of gravity, some of them are harmful. Our ability to function upon return to Earth will be hampered if these changes are not mitigated by “countermeasures.” The most obvious and most successful countermeasure demonstrated on Skylab was exercise. It minimizes the loss of muscular mass, strength, and endurance which weightlessness produces, and also stresses and protects the cardiovascular system just as “aerobic” exercise does on Earth. It was clear that exercise was going to be a mandatory part of long-duration space explorers’ days.

Another class of changes occurred but appeared to be self-limited:

- motion sickness, until the vestibular system adapts (in less than a week);
- loss of fluid from the blood and extracellular fluid compartments (complete in about 3 days);
- reduction in the number of circulating red cells (complete in about a month).

A third class of changes did not respond to Skylab’s few and primitive countermeasures. Their long-term threats to health remain to be assessed and counteracted:

- loss of bone mineral mass and strength;
- changes in function of the immune system;
- life-shortening effects of damage from space radiation;
- “cabin fever”—effects of the psychological stress of isolation and confinement. Skylab crews did not have a problem with this, but it remains a threat.

Work is being done today by NASA and the National Space Biomedical Research Institute (NSBRI) to develop countermeasures and test them on the International Space Station (ISS). Aside from the effects of weightlessness and confinement in a closed-up spacecraft, any time a group of humans lives in isolation for several months, illness or injury may occur. The practice of medicine in space poses some challenges:

1. Crew size is small. There may or may not be a physician on board. There certainly will not be a well-staffed operating room or even an emergency room. Trained people and the equipment they need are the primary shortages in space.
 2. Weightlessness itself may confound diagnosis and complicate treatment.
 3. Help is far away.
- How can these medical challenges be met and the risks minimized? For Skylab, the first tool was crew selection. Only people in excellent health and with few or no risk factors for disease were qualified for spaceflight in those days. And this method succeeded; no serious illness or injury occurred. The second was training. A physician flew only on the first crew, but two members of each

crew were well trained in basic methods of diagnosis and treatment so that, under the direction of a flight surgeon in Houston, they could examine a crewmate and carry out some treatment procedures. The equipment on board approximated that in the office of a general practitioner, plus a dental kit that allowed procedures up to and including extraction. Finally, for serious problems, return to Earth was available. Skylab had an “ambulance,” the Apollo Command/Service Module, docked and ready to return an ill crew member.

Human Factors and Design. Skylab was built and flown in the dawn of human spaceflight, and it was a challenge posing and answering all the questions about how to design a hotel-cum-laboratory in which the curse and blessing of gravity were absent, and everything were afloat (Fig. 8). Questions, options and experiments were everywhere. Should the workspace be cylindrical, following the shape of its enclosure, with equipment around the outer walls, or should floors and ceilings be installed? (Skylab tried both.) How would people move from place to place—head first or belly button first? (The latter was preferred.) Were chairs needed? (No!) What sort of restraints worked best to attach you to where you were working, eating, or sleeping? And what is the best arrangement for going to the bathroom?



Figure 8. Astronaut Edward Gibson is shown “floating” next to the dinner table on Skylab. The third Skylab crew celebrated Thanksgiving in 1973 by having their turkey dinner at this table. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

All of these “human factors” questions were answered and documented beautifully in a series of “Skylab Experience Bulletins” that defined every solution, from switch shapes to sleep stations. Space Station designers could proceed with confidence.

One of Skylab’s major achievements was unplanned—it demonstrated the practicality of external maintenance of a spacecraft by pressure-suited humans. Suits and an airlock were provided, but the only planned task was replacing film in some of the solar cameras. But Skylab’s launch problems created an opportunity to do much more. Erection of the stuck solar panel was successful, and subsequent trips outside included erecting a better solar shade for the workshop and repairs to camera doors and stuck relays. These successes helped to remove the aura of danger from spacewalks and has led to their extensive use in the Shuttle and ISS.

Knottier problems were those of maintaining a livable, comfortable environment in a completely closed space for a long time. Pressure, temperature, oxygen, and carbon dioxide content, humidity and contamination all had to be controlled. Skylab used a “brute force” approach to these problems. Its life support system was “open-loop”—all of the oxygen, nitrogen, clothing, water, and food needed for the three missions was launched with the workshop. Activated charcoal filters were used to control odor and contaminants and were replaced at intervals. Carbon dioxide was removed from the atmosphere by a “molecular sieve”—it was trapped in the interstices of lava-like mineral beds, then dumped overboard as the beds were periodically exposed to vacuum. There was no attempt to reclaim oxygen or water from waste or to grow food. The system was not designed to be replenished or repaired—it was a prototype. This fact would hamper plans for revisiting Skylab after the Space Shuttle became available.

The system worked beautifully. However, on a trip to Mars, a system that does not recycle at least water will be prohibitively heavy. NASA has been working ever since on “closed-loop” life support systems.

Science and Human Productivity. The broad question asked of Skylab was this: could humans remain physically well in space and also function effectively as scientists? The broad answer was “yes.” Weightlessness does not affect cognitive function in any unique way. The environment has its challenges; equipment needs to be designed with weightlessness in mind; any forgotten spares or tools will just have to be done without. But these constraints have been experienced—and overcome—by other expeditions, from Captain Bligh’s voyage to the present scientists at the South Pole.

The principal problem is the scarcity of people in orbit and hours in the day. The crew had to run the hotel, cook their own meals, and then do the experiments. Also, physical tasks learned on Earth must be relearned in weightlessness. The first trial of a complex task takes more than twice as long in space; three or four repetitions are required to regain speed.

Nonetheless, the three Skylab crews became quite efficient, aided by the relative simplicity and good human engineering of the Skylab design. An average of 7.5 hours per person per day was used for science. The results were all that the investigators had hoped for:

- The medical results form the most complete and most accurate physiological data set ever collected in space. It has still not been equaled either by Russian or American flights.
- Astronomers published from the Skylab solar data and images for more than 20 years. Among the advances were the most complete and accurate measurements of solar flares, a huge increase in the observations of the solar corona, and unique records of the comet Kohoutek.
- Multispectral, targeted image sets of land, ocean, and meteorological features highlighted the Skylab Earth Observations program and stimulated a much better understanding of weather and climate. In addition, the crew members were privileged to see breathtaking views of their home planet (Fig. 9).

In summary, human-operated spacecraft were proven useful as scientific platforms, both as vantage points for Earth and astronomical observations above the atmosphere and as laboratories providing microgravity to physical and life scientists (5).

Epilogue: Skylab Returns to Earth, 1979

The final lesson of America's first Space Station was that we had better plan more effectively how to dispose of our space assets. When the third mission to Skylab launched in November 1973, discussions of its future had already begun. It was considered possible that the Orbital Workshop could be revisited and even perhaps refurbished using the Space Shuttle then in development and due to be launched in 1979. That decision could wait, provided that after its final mission, Skylab was parked in a high enough orbit to remain aloft. The one thing to be avoided was a gradual decay of its orbit that resulted in uncontrolled reentry; there were some structural elements of Skylab that would survive reentry and could pose a threat to people and structures on Earth.

Of course, that is what happened, and here is why. The plan had been for the third crew, at the conclusion of their mission, to fire the Command and Service Module's (CSM's) 20,000-pound Service Propulsion System engine to raise the orbit of Skylab substantially—high enough to ensure an orbital lifetime of at least 10 years. But Rockwell engineers warned that the docking apparatus used to connect the CSM to the Workshop was too fragile and might sustain damage during the burn—might even risk damage to the CSM and endanger the crew. An altitude boost using the small reaction control system jets would raise the altitude sufficiently to keep Skylab in orbit until 1981, said the experts—2 years after the initial flight of the Shuttle. So it was decided and accomplished.

This careful plan was then dismantled by two factors. The first was a 2-year slip in the Shuttle development schedule; the first launch was delayed until 1981. The second was the most active solar cycle yet recorded. As solar activity increased toward its 11-year peak in 1980, it heated the Earth's upper atmosphere, increasing drag at Skylab's altitude and hastening its decay. By 1978, it had become clear that Skylab was not going to make it. Now the only thing to do was to try to steer it to a safe reentry point.

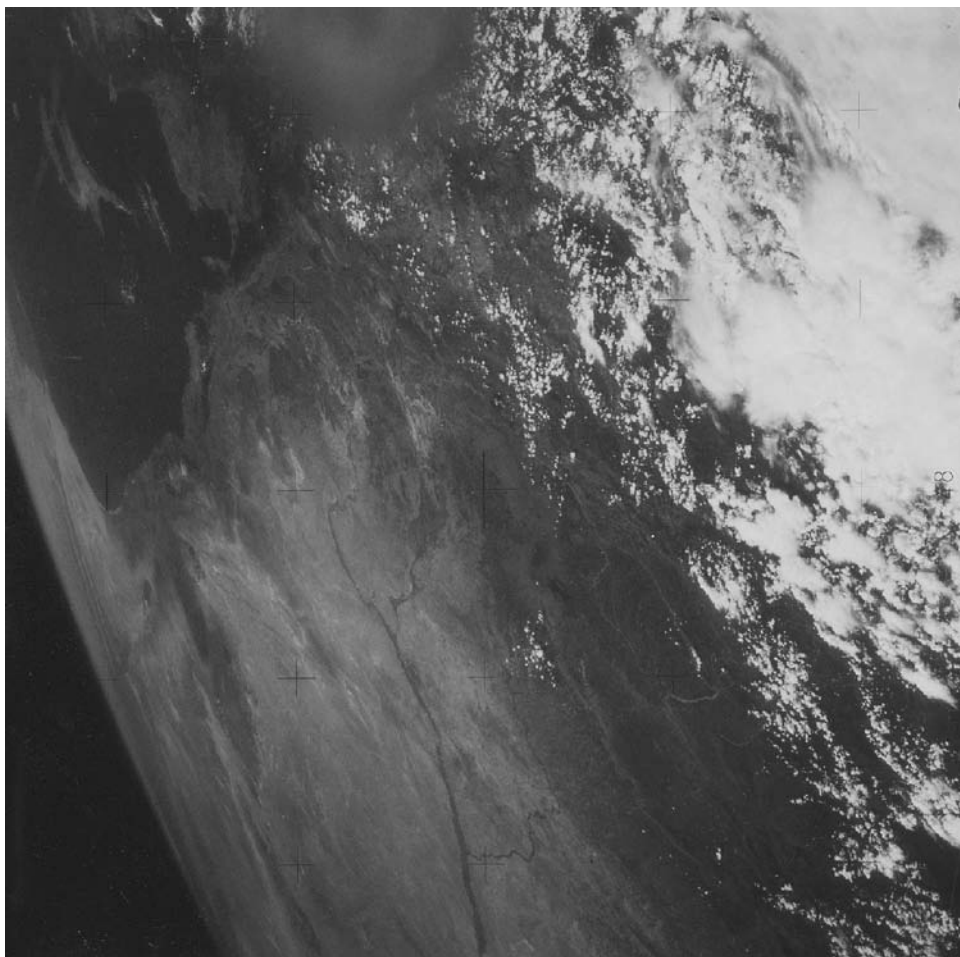


Figure 9. This picture was taken when Skylab was in orbit over eastern Turkey looking south. The eastern shore of the Mediterranean Sea is at the top center of the picture. The island of Cyprus is at the upper right, and the Nile River Delta is directly to the south (i.e., above) of Cyprus. The Euphrates River can be clearly seen in the center of the picture flowing east through Iraq. The Dead Sea, the Gulf of Aqaba, and the triangular Sinai Peninsula are visible at the top of the picture. In this one photograph, the Skylab crew was able to capture this view of the cradle of civilization. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The team assembled to monitor and attempt to control Skylab's reentry did not have much to work with. Skylab had no propulsion system. All it had was an attitude control system consisting of control moment gyros, which had exceeded their lifetime and were failing, backed up by cold-gas jets whose nitrogen fuel was almost exhausted. So, very carefully, this team commanded Skylab to rotate in and out of its "maximum drag" orientation, hoping to bring it to reentry altitude at just the right longitude to force it down over the Indian Ocean.

The reentry team did a remarkable job. They lost attitude control of Skylab over the North Atlantic on the fringes of the atmosphere. It continued over

southern Africa, and finally broke up over the Indian Ocean west of Australia. Most of the debris fell into the ocean, but a little of it landed in southwestern Australia, east and south of Perth. Fortunately, this area is sparsely populated, and there was no injury or property damage. Many Australians found and returned pieces of Skylab to NASA for analysis of the effects of reentry. At least one proud Aussie had an intact, charred Skylab oxygen tank to hang over the front door of his pub. If you ever see it, reflect on how space flight has united the citizens of the world.

BIBLIOGRAPHY

1. Oberth, H. *Die Rakete zu den Planetenraumen* (The Rocket into Planetary Space), Oldenburg, Munich, 1923, reprinted by Uni-Verlag, Nurnberg, 1960 and *Wege zur Raumschiffahrt*, 1929, reprinted by Kriterion, Bucharest, 1974.
2. von Braun, W., F.L. Whipple, J. Kaplan, H. Haber, W. Ley, and C. Ryan. Man will conquer space soon, *Collier's Magazine*. A series of seven articles starting on March 22, 1952 and ending April 30, 1954.
3. Logsdon, J.M. (ed.). *Exploring the Unknown, Vol. I, Organizing for Exploration*. NASA SP-4407, U.S. Government Printing Office, Washington, DC, 1995, See pp. 500–548 for mention of possible missions that later became Skylab.
4. Belew, L.F. (ed.). *Skylab, Our First Space Station*. NASA SP-400, U.S. Government Printing Office, Washington, DC, 1977.
5. Belew, L.F., and E. Stuhlinger. *Skylab, A Guidebook*. NASA EP-107. This document can be found at: <http://history.nasa.gov/EP-107/contents.htm>.

JOSEPH KERWIN
Houston, Texas

SOLID FUEL ROCKETS

Solid Fuel Rocket Fundamentals

A solid fuel rocket is distinguished from a liquid fuel rocket by the type of fuel that it uses. It is more accurate to refer to the two basic types of rockets as solid propellant and liquid propellant rockets. Both types of rockets generate thrust by exhausting combustion products through a supersonic nozzle. Rocket propellants contain both a fuel and an oxidizer to produce energy. Both carry their own oxygen, so they can operate above the atmosphere and can be used in space.

A solid fuel rocket, commonly called a solid rocket motor, is a completely self-contained device that converts chemical energy into kinetic energy in a controlled way. Although there is much to know about the science, engineering, and manufacture of solid rockets, they are simple devices. Solid rockets consist of four main components: (1) propellant grain, (2) a case that is the thrust chamber that contains the pressurized combustion gases, (3) a nozzle for directing and accelerating the gases away from the motor, and (4) an igniter.

The propellant grain consists of the propellant charge shaped to deliver the desired thrust profile. The case contains the pressure of propellant combustion and is frequently a major portion of the vehicle airframe. Insulation is necessary to protect the case from the high-temperature combustion products. The igniter provides the heat necessary to initiate combustion at the propellant surface. The nozzle directs and accelerates the propellant exhaust gases.

The propellant is typically a solid rubberlike material, similar to a pencil eraser, that contains fuel and oxidizer particles. The propellant grain can have a wide variety of shapes, but it is generally a hollow cylinder designed to burn from the inside out, thereby not exposing the case to the extreme temperatures until near the end of the burn. The propellant is bonded to the case.

The rate at which the propellant burns and thereby generates thrust is designed into each propellant formulation. For instance, propellants that have fine oxidizer particles burn at higher rates than formulations that contain coarse oxidizer particles. Burn rate also varies with combustion chamber pressure, initial temperature of the propellant grain, and other factors.

The internal shape of the grain, which can vary the amount of exposed surface area, hence the burn rate, is typically established when it is cast (poured) and cured inside the case. An igniter, generally located in the head end and fires down the bore or grain center perforation, initiates the motor.

Hot combustion gases are ducted from the motor through a supersonic nozzle, which accelerates the gas and converts the pressure and temperature of propellant combustion to kinetic energy. Directional control is attained through a number of different means, but commonly the nozzle is jointed, allowing it to be vectored by mechanical actuators. Figure 1 illustrates a typical solid rocket booster.

History. Liquid fuel rockets date back only to Robert Goddard in the 1930s; solid fuel rockets have been around since the thirteenth century when the Chinese invented them. They were used in everything from fireworks to warfare. Modern solid fuel rocket history dates to the mid-1950s when the U.S. military started experimenting with a sulfur-based sealant made by Thiokol. This sealant

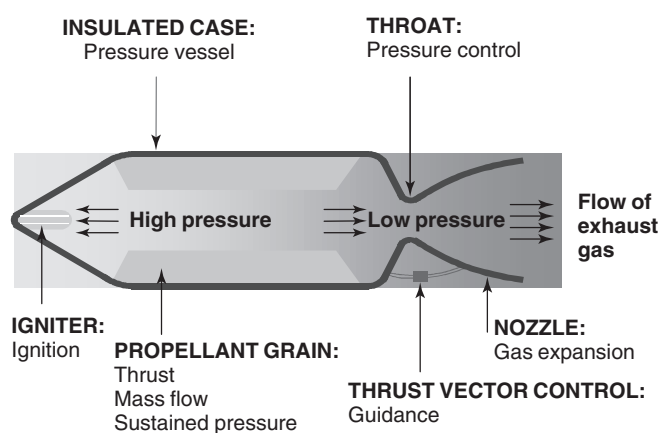


Figure 1. A typical solid rocket booster.

and various solid fuels and oxidizers were used to make solid propellants. An early result of this work was the Sergeant missile. A scaled-down version of the Sergeant was clustered as the upper two stages of the Jupiter C (eleven Sergeants as the second stage and three as the third stage) that put the first U.S. satellite, Explorer 1, into orbit on 31 January 1958.

As the Cold War escalated, it became clear that solid propulsion offered several advantages for ballistic missiles: low cost, high reliability, low maintenance, long-term storability, packaging efficiency, and quick response. This led to a major development of solid fuel rocket capability in the United States and resulted in a long line of silo- and submarine-based ballistic missiles such as Minuteman, Peacekeeper, and the Poseidon/Trident family that have long-range capability.

In the 1960s with an eye to space access, two very large solid rockets were tested. One was 156 inches in diameter, and the other was an enormous 260 inches in diameter. These tests demonstrated, very early, the capability to make large, high-thrust solid rockets. This development led to large segmented solid booster systems for both the Titan and the Space Shuttle launch vehicles. Many space launch vehicles benefit from additional energy and thrust gained by strapping solid rockets around a liquid rocket core vehicle.

Because of their many advantages and inherent design flexibility, solid rockets are used in a wide range of applications besides launching. They have been used to transfer spacecraft from low Earth orbit to higher orbits, including geosynchronous altitudes. They have also been used as retrograde motors for stage separation and planetary descent. Solid rockets were used to reduce the descent velocity of the Sojourner spacecraft as it approached the Martian surface.

Many military tactical propulsion devices use solid rockets. These have a wide range of applications, including surface-to-surface, air-to-air, air-to-surface, and surface-to-air. These propulsion systems start with a diameter of approximately two inches and range to much larger, longer range systems. For example, solids power the Patriot antimissile system used successfully in the Gulf War in 1991.

Principle of Operation. Solid rockets produce thrust by the same principle as a child's balloon. Gases that exit the balloon in one direction push it in the opposite direction, illustrating Newton's famous theorem of "equal and opposite reaction." Modern rockets take the principle of the balloon to a very high level of refinement (1). In a solid rocket, the gases are generated by burning a high-energy propellant. Unlike a balloon, solid rocket pressure vessels are rigid and allow much higher internal pressures. Modern rockets operate at 2000 psi, and chamber temperatures are 5000°F. The main component of thrust can be calculated as the product of the propellant mass flow rate and the velocity of the gases as they exit the nozzle. Under steady-state operating conditions, the propellant burning surface area, density, and burn rate determine the mass flow rate. Nozzle exit velocity is determined by propellant formulation, chamber pressure, nozzle geometry, and ambient pressure (2).

Manufacturing Flow. The type of motor case dictates the initial portion of the manufacturing flow. Solid rocket cases are of two types: filament-wound composite and metal. For filament-wound composite cases, the insulation is placed

over a removable mandrel, cured, and machined, and the filament-resin mixture is wound over it. After the case is cured, the mandrel is removed. In a metal case motor, the case is fabricated first, and the insulation is laid up inside the case and then cured. This curing process also bonds the insulator to the case.

An elastomeric material, called a liner, is then applied to the inside of the insulated case to provide a bonding agent between the insulation and the propellant. A removable core is placed inside the case to provide the initial surface geometry for the propellant.

Propellant ingredients are mixed in bowls (up to 1800 gallons in size) similar to those used for making bread dough. When mixed, the propellant has the consistency of runny peanut butter. The propellant is then poured into the motor around the core and allowed to cure (become solid). Once the propellant has cured, the core tooling is removed. The loaded case is then ready for final assembly. In final assembly, the nozzle, igniter, and other mission-unique hardware such as cable raceways, thrust vector actuation systems, and interstages for connecting to payloads or other rocket stages are installed.

Propellants

A solid rocket propellant consists of an elastomeric polymer that may be filled with as much as 91% of solid particles. The solid propellant may have 4 to 12 ingredients and is formulated using a combination of a fuel and an oxidizer that are intimately mixed on a microscopic level, and in some cases, are parts of the same molecule. Initially, the ingredients that make up the solid propellant are mixed together to form a thick liquid containing suspended solid particles. Once the propellant has been cast into the motor case, the mixture hardens enough to maintain its mechanical integrity and retains sufficient elasticity to prevent cracking as it experiences induced stresses from operation (high pressure) and storage (thermal expansion and contraction) (3).

Parameters for Formulating Propellants. Many interacting factors must be considered when formulating a solid rocket propellant. A primary consideration is for the propellant to provide sufficient energy to meet mission requirements. In addition to specific impulse (I_{sp}), a measure of the amount of thrust the propellant provides per unit mass, solid rocket motors are often volume-limited. Hence propellant density must be considered. The operating pressure and mass flow rate of a solid rocket motor depend on the ballistic characteristics of the propellant, so these characteristics must be tightly controlled. Propellants typically use very energetic materials to achieve the performance objectives of the rocket motor. Energetic materials are often susceptible to ignition stimuli. They may be shock-sensitive, toxic, or apt to react explosively when damaged and ignited. Each of these factors must be considered when evaluating a combination of propellant materials.

There are other issues to be considered when formulating propellants, including their cost, maturity, reproducibility, manufacturability, compatibility, and service life. Finally, because rocket motors are generally not used immediately after they are manufactured, the propellant must maintain its properties over time and under the environmental conditions required for the given application.

Propellant Categories. A variety of different types of solid rocket propellants have been developed and deployed during the past several decades. Among other things, they can be grouped by signature, hazards, or application. The propellant exhaust signature is of major importance in military applications. Exhaust signature in this context refers to the visible plume that is emitted from the rocket booster when it is operating; the plume allows the adversary to track a vehicle visually back to its launch site. As a result, formulations are often designated as metallized (smoky), reduced smoke, or minimum smoke propellants. Reduced smoke propellants have very little primary smoke (smoke produced within the combustion chamber) but may form considerable secondary smoke (similar to the contrail which sometimes forms behind high-altitude aircraft) under certain atmospheric conditions. Minimum smoke propellants produce a minimum amount of both primary and secondary smoke. They are formulated to produce only gaseous combustion products.

Propellants may also be categorized by their potential for unplanned ignition. Those that have an explosion hazard are given a Class 1.1 designation. Those that have a mass fire hazard but not an explosion hazard are termed Class 1.3 propellants. Class 1.3 propellants are usually formulated using ammonium perchlorate (AP) as the principal oxidizer; Class 1.1 propellants often incorporate large quantities of the detonable nitramines HMX and RDX, and/or energetic nitrate esters such as nitroglycerin (NG) or butane triol trinitrate (BTN).

Propellant Composition. A number of different materials are required in solid propellants to provide the desired properties. Table 1 gives a list of the categories of ingredients in solid propellants and the most common examples of each, along with brief comments. Not all propellants require ingredients from each category. By far the most common oxidizer in rocket propellants today is ammonium perchlorate (AP). It provides high energy, high density, excess oxygen, low detonability, low cost, ballistic tailorability, good mechanical properties, and good aging characteristics. Nitrates generally suffer from insufficient energy, poor ballistic characteristics, and moisture sensitivity. Ammonium nitrate (AN) is the cheapest oxidizer available for use in solid propellants, but its other drawbacks have prevented its widespread use. Newly developed oxidizers have not yet been extensively used due to their higher cost, immaturity, or limited availability.

Binders that provide mechanical integrity to propellants include polybutadienes, polyethers, and polyesters. In most cases, the raw materials are low molecular weight polymers that contain reactive functional groups. They wet and suspend the solid materials and then cross-link (using an appropriate curative). Energetic plasticizers are typically used in combination with oxygenated binders. They provide both increased energy and improved mechanical properties. High-energy propellants contain as much as three times more plasticizer than polymer.

Metal fuels are incorporated into solid rocket propellants to provide increased specific impulse (I_{sp}) and density. I_{sp} is a measure of the theoretical performance of a propellant and is defined as the calculated specific impulse at a chamber pressure of 1000 psi expanded to 1 atmosphere with an optimum nozzle expansion ratio. Beryllium is the metal that provides the highest I_{sp} . However, it has toxicity problems that limit its application. Boron and magnesium have been used in some systems, but aluminum is by far the most common metal fuel. Its

Table 1. Common Propellant Materials and their Functions

Functional category	Common examples	Comments
Solid oxidizers	Ammonium perchlorate (AP), other perchlorates, ammonium nitrate (AN), other nitrates, ammonium dinitramide (ADN), hydrazinium nitroformate (HNF)	AP is used in all but minimum smoke propellants. AN is low cost but has numerous drawbacks, including moisture sensitivity, little ballistic tailorability, phase transitions, and aging concerns. ADN, HNF are still immature in U.S.
Energetic monopropellants	Nitramines: Cyclotrimethylenetrinitramine (RDX), cycloctetramethylenetetraminetrinitramine (HMX), hexanitrohexaazaisowurtzitane (CL-20)	Nitramines are used in most Class 1.1 and some Class 1.3 propellants. Provide increased I_{sp} . These three nitramines provide similar I_{sp} in aluminumized propellants. CL-20 is the densest and most oxygen-rich, but least mature.
Binders	Hydroxyl-terminated polybutadiene (HTPB), carboxyl-terminated polybutadiene (CTPB), polybutadiene acrylonitrile (PBAN), polyethylene glycol (PEG), polypropylene glycol (PPG), nitrocellulose (NC), glycidyl azide polymer (GAP)	HTPB, CTPB, PBAN are the most common Class 1.3 propellant binders. PEG, PPG, NC are generally plasticized with nitrate esters. NC and GAP are energetic.
Curatives	Isocyanates, epoxides	Isocyanates are used to cure hydroxyl-terminated polymers; epoxides cure PBAN, CTPB
Fuels	Beryllium, aluminum, magnesium	Beryllium gives the highest I_{sp} , but is toxic. Aluminum has better I_{sp} and density than magnesium but is not as easily ignited.
Plasticizers	Diethyl adipate (DOA), dioctyl phthalate (DOP), triacetin, other inert esters, nitroglycerin (NG), butanetriol trinitrate (BTTN), trimethylolethane trinitrate (TMETN), other nitrate esters	DOA, DOP, and similar esters are used with nonpolar binders such as HTPB. Triacetin is used as a desensitizer for nitrate ester propellants. Nitrate esters provide increased I_{sp} . NG is highest density and highest performance nitrate ester. Other nitrate esters are used to decrease detonability (compared with NG) or to give better low-temperature properties.
Stabilizers	AO2246, <i>p</i> - <i>n</i> -methyl nitroaniline (MNA), nitrodiphenylamine (NDPA)	AO2246 prevents oxidative cross-linking of HTPB; MNA, and NDPA stabilize nitrate esters
Ballistic modifiers	Iron oxide, aluminum oxide, oxamide	Iron oxide; aluminum oxide accelerates burn rate; oxamide and other coolants slow burn rate.

use provides a theoretical I_{sp} density improvement of about 10% compared to a nonmetallized formulation.

Polymers, plasticizers, fuels, and curatives constitute the bulk of most solid propellant formulations, but several other ingredients are used to tailor one or more of their properties. These ingredients include ballistic modifiers such as iron oxide (which increases the burn rate), combustion stabilizers, chemical stabilizers (for nitrate esters and some polymers), processing aids, cure catalysts, and bonding agents. Bonding agents enhance the bond between oxidizer particles and the binder and greatly improve mechanical properties.

Ballistic Properties. The ballistic (or combustion) characteristics of a propellant include burn rate as a function of pressure and temperature, combustion stability, and the completeness of combustion. Propellants that exhibit large changes in burn rate as a function of pressure are not often used. The dependence of propellant ballistics on temperature is particularly important for tactical rockets, which must operate over a wide temperature range (-50°F to 150°F).

Ballistic properties can be controlled by a number of formulation variables. Among the most significant variables that affect burn rate are AP particle size (smaller = faster burn rate), total solids loading (higher = faster burn rate), the presence of ballistic modifiers (can increase or decrease burn rate, depending on the modifier), the curative used, and the metal content and particle size.

Propellant Case Bond. One other factor that must always be considered when developing rocket motors is the means for attaching the propellant to the case wall. A thin layer of adhesive, or liner, is usually applied to the motor case before manufacturing the propellant. The liner is usually cured to some degree before the propellant is cast and then is cured more completely along with the propellant. The liner is quite significant because it must bond to both the inner case wall, which may be insulation or case material, and to the propellant. Unless it is to function also as a barrier to the diffusion of propellant ingredients such as plasticizers or curatives, it is usually formulated as an elastomeric material that has a backbone chemically similar to that of the propellant.

Propellant Grain Design

System constraints and requirements such as allowable length, volume, maximum pressure, and thrust profile largely drive the grain design. Additional factors such as clearance for the nozzle, the thrust vector control (TVC) method, clearance for the igniter, and location of the ignition system are also important considerations.

Design Considerations. Figure 2 shows an example of some features used in grain design. Grain structural loads often require stress relief features such as flaps or slots. Features such as radial or longitudinal slots are incorporated into the propellant grain to obtain the desired motor performance. Grain structural requirements drive such features as allowable web fraction (grain thickness/available distance) and the presence or absence of flaps and stress relief slots. A flap is an elastomeric piece bonded to the grain that disconnects the grain from the case. Flaps typically do not have a large impact on the ballistic performance of the motor; however, stress relief slots can present significant design complications. Stress relief slots are slots located in the grain solely for structural reasons.

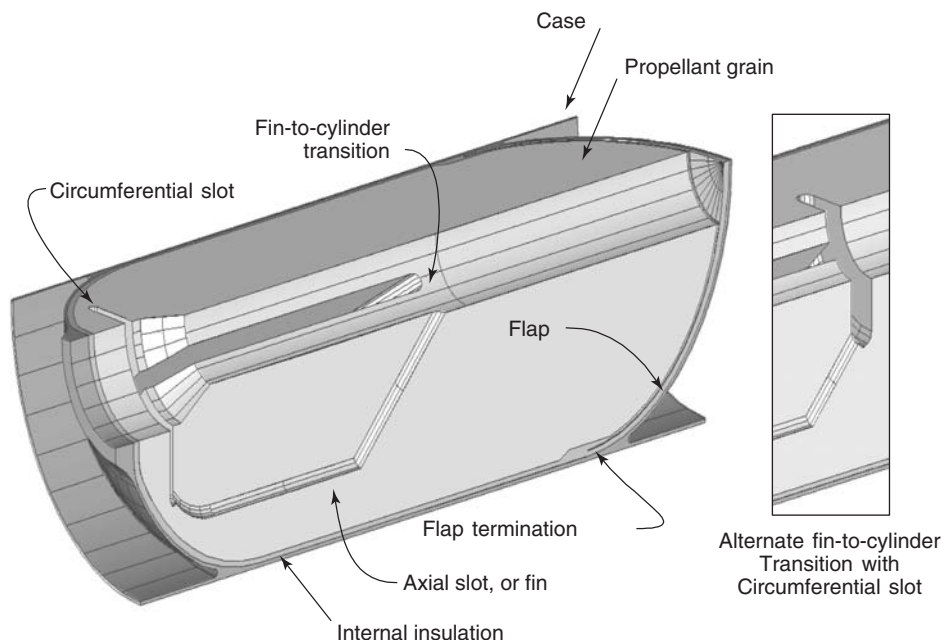


Figure 2. Features used in grain design.

Structural analysis of the solid propellant grain is important because structural failure of the propellant can result in catastrophic failure of the motor. A crack in the propellant or a separation in the propellant-to-case bond will increase the surface area, causing increased pressure, which may burst the case. A crack or bond separation may also allow flame to reach the case sooner than planned, resulting in case burn-through. Structural loads that are likely to cause propellant cracking or bond separation include cure and thermal shrinkage, internal pressurization, and acceleration.

In designing a motor, the stresses and strains induced in the propellant by these loads must be kept below the propellant's capability. Most methods for reducing grain stresses and strains reduce the amount of propellant in the motor, so careful consideration must be given to both propellant loading and structural integrity. When designing a propellant grain, it is assumed that propellant burns normal to the grain surface. The burning surface area as a function of distance burned is then used to determine the propellant mass flow rate as a function of time.

Grain Manufacture. Typically, the grain forming tooling is placed in the motor case, and the propellant is cast between the case and the casting tooling. The propellant is then cured at elevated temperature, and when cure is complete, the tooling is removed from the motor. The two primary methods for introducing the propellant into the motor case are (1) pressure casting and (2) vacuum casting. In pressure casting, the propellant is forced through a tube into the motor. In vacuum casting, a vacuum is introduced in the interior of the motor case, and the propellant is pulled into the motor. Casting tooling may be removable; the tooling is removed from the motor after propellant cure is complete or may be left in place. Sometimes both types of grain forming tooling are used in the same motor.

Structural requirements may require curing the propellant under pressure. The combination of vacuum casting and pressure cure results in grains that have few defects such as voids. Grains may be cast with simple tooling, and more complicated slots are machined into the grain after propellant cure. This results in an additional manufacturing step, but changing the grain design is as simple as altering the machining program, rather than altering expensive hard tooling.

Another type of grain fabrication is extrusion. The propellant is forced through a die and cut to length. This process is not typically used for large rocket motors due to the difficulty of retaining the grain in the rocket motor case.

Motor Case

The solid rocket motor case is the pressure vessel that contains the solid propellant and provides a structural interface to external components. The case is designed to contain the high pressures generated by the burning propellant during motor operation. It is also the airframe that transfers thrust from the motor to the launch vehicle or missile system. As such, the case must incorporate features for attaching the pressure vessel to other stages, payloads, launch support equipment, or other support structures. Basic loads during motor operation are shown in Fig. 3.

Metal or composite materials are generally used in case design. Metals are generally less costly, more damage tolerant, and better characterized than composites. Composites generally weigh less due to the high strength to weight ratio but are less stiff for a comparable thickness. A composite case can be as much as five times lighter than a metal case. This weight reduction requires less propellant for equivalent vehicle performance, which ultimately can lead to a less expensive motor.

Insulation

Internal insulation is the heat barrier between the case and the propellant. It protects the case from reaching temperatures that would endanger its structural integrity. Insulation also serves to (1) buffer the transmission of case stresses to the propellant, (2) inhibit burning on designated propellant surfaces, (3) provide a pressure seal for the case, and (4) limit the diffusion of chemical components to or from the propellant. Insulation may also be located on the exterior of the rocket motor to protect the case from aeroheating and to provide damage tolerance and protection from the elements.

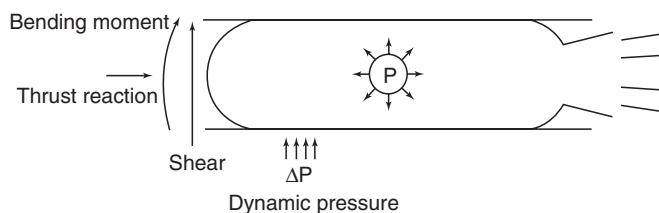


Figure 3. Basic loads during motor operation.

Insulator materials are organic compounds that consist of reinforcing fillers contained in a binder. Fillers contribute to char-layer strength and include silica, asbestos, carbon or Kevlar fibers, nylon, and glass cloth. Elastomers and plastics are two classes of binders.

Internal insulator design predominantly involves provision for material char and ablation. The general equation for computing design thickness is

$$\text{Thickness} = (ET \times \text{MAR}) \times \text{SF} + \text{TP} + \text{MT}$$

where

ET = exposure time

MAR = material affected or char rate

SF = safety factor

TP = thermal protection thickness

MT = manufacturing tolerance

Nozzle

The functions of a nozzle are to provide a ballistic throat for the motor, control motor pressure, direct subsonic gases into the throat, additional thrust, and thrust vector control (TVC). Solid rocket motors use ablative nozzles. Nozzle design requires knowledge of gas dynamics, heat transfer, combined thermal-structural loads, and material science. The nozzle environment is harsh; temperatures can reach 6000°F. Pressures can exceed 3500 psi, and exposure time can reach 300 seconds. Temperatures, pressures, and gas velocities vary greatly from the inlet region of a nozzle through the throat and into the exit cone.

Typical Nozzle Designs. Nozzle configurations are either external or submerged. The external nozzle extends aft beyond the nozzle-to-chamber interface. The submerged nozzle extends forward into the case. Submerged nozzles provide a separated flow region in the aft motor that results in smoother gas flow into the throat region and decreased erosion of the motor insulator. Length-constrained motors employ nozzles submerged 10–40%.

Function. The nozzle inlet directs subsonic gases smoothly into the throat where they transit to sonic flow. The throat controls motor pressure by the initial size and material erosion rate interactively with the propellant burn rate. The exit cone expands and controls the supersonic gases. Contoured exit cones provide efficient and optimum expansion or thrust in the shortest length (4,5). Nozzles provide motor pressure control, efficient gas flow, and thermal protection for primary structural components by erosion and thermal degradation of ablative insulators.

Propellant exhaust gases heat up the nozzle surface of the ablative material via radiation and convection. Energy from the surface is conducted into the ablative material at rates dependent on the ablative reinforcing fiber type, the binder material (resin system), and the phenolic resins, or elastomeric binders. Heat conducted into the ablative material causes resin or binder degradation

resulting in pyrolytic gases that percolate through the porous char layer and provide internal “cooling.” Surface erosion is the result of excessive material temperatures (i.e., melting), mechanical erosion, or chemical reaction between the ablative constituents and the exhaust gases (6,7).

Nozzle materials. Materials consist of structural or ablative/insulators. Structural materials include metals fibers-reinforced phenolic or epoxy composites. Ablative selection requires knowledge of gas pressures, temperatures, velocities, and propellant exhaust characteristics. The properties of commonly used ablative or insulators, as well as some throat materials, can be found in Reference 4.

Carbon or graphite ablatives have been mostly derived from cellulosic or polyacrylonitrile (PAN) fiber precursors, subsequently carbonized or graphitized, woven into fabrics, and preimpregnated with phenolic resins. Silica and glass phenolics consist of melted glass filaments subsequently combined into fiber strands woven into fabrics and impregnated with phenolic resin.

Molding compounds consist of glass, silica, or carbon fibers or woven fabrics chopped into short lengths and mixed into phenolic resin. Elastomeric-based ablators or insulators combine chopped fibers or other fillers into the basic Buna-N, silicone, or ethylene-propylene diene monomer (EPDM) rubber compound. Most ablatives used in the inlet region must conform to structural component deformations and match the motor aft insulator erosion characteristics. They are typically Buna-N, silicone, or EPDM-based elastomeric material reinforced with short chopped glass, carbon, or Kevlar fibers. External nozzles often incorporate glass, silica, or carbon-reinforced phenolics because they provide better erosion resistance than the elastomeric-based materials.

Throat material selection is based on motor performance requirements and/or type of propellant. A high-performance motor with aluminized propellant often requires a low eroding throat material. Typical materials range from carbon phenolic used on the large Space Shuttle nozzle throat (8) to low eroding carbon-carbon or noneroding tungsten, which is used on high-performance tactical motors.

Exit cone ablatives require more erosion resistant material than inlets and typically incorporate silica-, carbon-, or graphite-reinforced phenolics. Reinforced elastomers do not have sufficient char strength to be used in the supersonic flow region of the exit cone where gas velocities or particle impingement tend to wash off low-strength charred materials.

Nozzle structures must withstand high loads due to combinations of motor pressures, ablative thermal expansion loads, thrust vector control (TVC) actuation, and external aerodynamic loads. Materials most often used are high-strength, aerospace-grade steel, aluminum, and titanium alloys.

Where weight is critical and stiffness is needed, polyacrylonitrile (PAN)-based reinforced epoxies are incorporated as structural components. These can be filament-wound, tape-wrapped, or hand laid-up.

Thrust Vector Control

Thrust vector control (TVC) uses external means (typically mechanical or fluidic) to alter the direction of the thrust, thus changing the direction of the missile's flight path. Figure 4 shows how one type of TVC, a movable nozzle, creates a

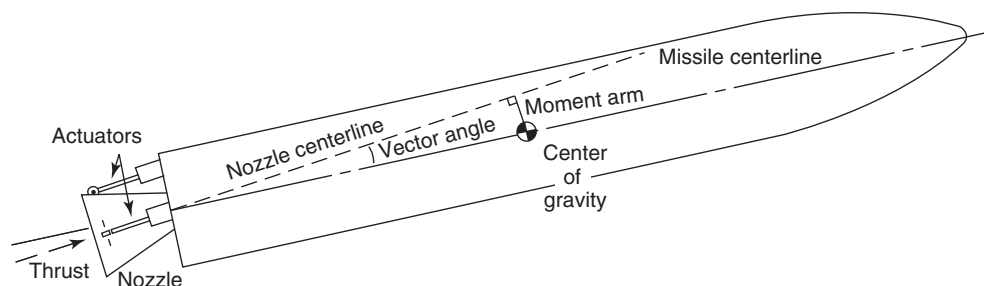


Figure 4. One type of TVC, a movable nozzle.

turning moment on the missile that is equal to the thrust times the moment arm. TVC systems can be configured to provide control in all three axes: pitch (up and down motion), yaw (side to side motion), and roll (rotation about the longitudinal axis of the missile). TVC can be thought of as the steering system or “rudder” of a missile (8–10).

Fixed-Nozzle TVC. Fixed-nozzle systems use thrusters, secondary fluid injection, or mechanical deflectors to change the thrust vector direction. Thrusters are multiple discrete nozzles placed in strategic locations on the missile to create the desired turning forces. The gas may be supplied by a pressurized bottle (cold gas), a gas generator (warm gas), or may be taken directly from the combustion chamber of the rocket motor (hot gas). Thrusters have very fast response time, and they also create a shock wave on the outside of the missile that can amplify the resulting force. Disadvantages include the expensive exotic materials required for hot gas valves and the weight and volume penalty of packaging. Thrusters are used on the fourth stages of Minuteman III and Peacekeeper.

In secondary injection TVC, a fluid (liquid or gas) is injected into the exit cone through the wall. This creates side forces from a combination of the thrust of the injectant jet and pressure imbalances from shock waves. The injectant can be an inert or reactive liquid, in which a chemical reaction results in additional side force. Gas injectants may come from gas generators (warm), or they may be bled directly from the motor combustion chamber (hot). Advantages include fast response and the thrust addition to the main flow. Disadvantages are the large packaging volume required and limited thrust deflection (about 6°). Liquid injection systems have been in production on the Titan III, Minuteman III, and Polaris. Gas injection systems suffer from material problems in the severe environment and have never reached production status.

TVC systems that use mechanical deflection include jet vanes, jet tabs, and jetevators. Jet vanes are aerodynamic fins inside the nozzle that rotate to provide pitch, yaw, and roll control. Roll control is one advantage of jet vanes. Another advantage is the low torque required to actuate a jet vane. Disadvantages include the exotic materials, required for the severe flow environment, the length penalty from packaging, and thrust losses due to the aerodynamic drag on the vanes. Jet vanes were used on the German V-2 missile and on the United States Sergeant, Talos, and Pershing missiles. They are currently used on the United States RIM-7 Seasparrow, the Vertical-Launch ASROC, and the AIM-9X.

Jet tabs are retractable surfaces mounted at the aft end of the exit cone that pivot in and out of the gas flow, creating shock waves and pressure imbalance. An advantage over jet vanes is that no thrust losses occur when the tabs are out of the flow. Jet tabs also have a low actuating force requirement. However, jet tabs do not provide any roll control, have high thrust losses, and also require exotic materials. Jet tabs were used on the MK-106 Tomahawk booster rocket motor, which was later replaced by the MK-111 motor with a movable nozzle.

The jetevator consists of a spherical ring mounted around the nozzle exit cone that can be rotated into the supersonic gas stream. This rotation creates a steering side force on the missile. Jetevators were operational on the Polaris, BOMARC, and SUBROC missiles. One advantage is a side force that is linear with the deflection angle. Disadvantages include large weight and volume, severe environment requiring exotic materials, and large thrust losses.

Movable-Nozzle TVC. In movable-nozzle systems, the nozzle is mechanically pivoted, which turns the hot supersonic flow of gases, thus changing the thrust vector. Movable nozzles are further subcategorized according to the location of the joint. If the entire nozzle and exit cone pivot as a unit, it is called a subsonic splitline. In the supersonic splitline, only the aft part of the exit cone pivots. Each has advantages and disadvantages; however, the supersonic splitline has never been in production due to manufacturing challenges.

Movable nozzles have lower thrust losses than the other types of TVC. However, a single movable nozzle cannot provide any roll control. Roll requires at least two nozzles. The Minuteman uses four hinged movable nozzles for pitch, yaw, and roll control, and the Space Shuttle has two booster motors with movable nozzles. Further categorization of movable nozzles is based on the type of joint. The flexible joint (also known as a flexible bearing or “flexbearing”) typically consists of alternating layers of an elastomeric material for flexibility and a rigid material (steel or composite) for strength and stiffness. The predictable and repeatable nature of the actuating force required to move a flexbearing nozzle is considered an advantage. Flexbearing nozzles are the most widely used TVC systems, as shown by application to the Space Shuttle boosters, Ariane, Trident, and Peacekeeper. The Space Shuttle RSRM flexbearing (Fig. 5) is the largest TVC system in production.

The ball-and-socket joint, also known as a trapped ball, has a spherical socket that rides inside a mating spherical ball surface. The ball-and-socket nozzle has certain advantages over the flexbearing, including higher vector angles, higher motor pressure capability, and less pivot point shift. Disadvantages include an unpredictable stick-slip friction force, susceptibility to contamination, and it requires an antirotational device to prevent the nozzle from “rolling” in the socket. Two Navy surface-launched motors that employ cold-trapped ball nozzles are the Tomahawk MK-111 booster motor and the Mk-72 Aegis ER booster motor.

The fluid bearing/rolling seal, also known by the patented name Techroll[®], is composed of a pair of rolling elastomeric convolutes that contain a fluid. The greatest advantage is the low actuating force required to move the bearing. The main disadvantages are the low structural stiffness and resulting large misalignment of the nozzle. This bearing is used on the Air Force Inertial Upper Stage (IUS) space motors.



Figure 5. Space Shuttle RSRM flexbearing—the largest TVC system in production.

The hinged movable nozzle is supported on thrust pins that ride in journal bearings. The hinged nozzle has the advantages of high deflection capability and low actuating force. The main disadvantage is that it provides control in only one direction (pitch or yaw) because it rotates only about one axis. The most well-known application of hinged nozzles is the Minuteman, which has four hinged nozzles on the first stage and thus provides all axes of control (pitch, yaw, and roll).

The gimbaled nozzle is an extension of the hinged nozzle; the thrust pins about which the nozzle pivots are themselves mounted in a rotating assembly that pivots about another set of thrust pins which are located 90° around the nozzle from the first set. This gives the nozzle omniaxial motion capability. This approach also has high deflection capability but requires a large envelope for packaging the gimbal mechanism. The gimbaled nozzle has been tested on the ground and in flight but has not been in production.

The rotatable nozzle is a canted nozzle mounted on a rolling bearing so that it can pivot about the motor centerline. Its main advantage is the low actuating force required. However, it is limited to motors that have multiple nozzles because movement of the nozzle induces pitch, yaw, and roll moments that must be balanced by the other nozzle. Another drawback is that the rotational deflection required is much larger than the actual vector angle achieved. This was an operational system for the Polaris missile.

Igniter

The purpose of the igniter is to provide the heat and pressure rapidly needed to start propellant combustion. Figure 6 shows a typical solid propellant rocket motor that has an axial flow igniter and some of the factors that affect motor ignition. These factors include igniter location, mass flow rate and action time of the igniter, impingement of the igniter output on the propellant surface, propellant grain geometry, combustion chamber free volume, and nozzle throat size.

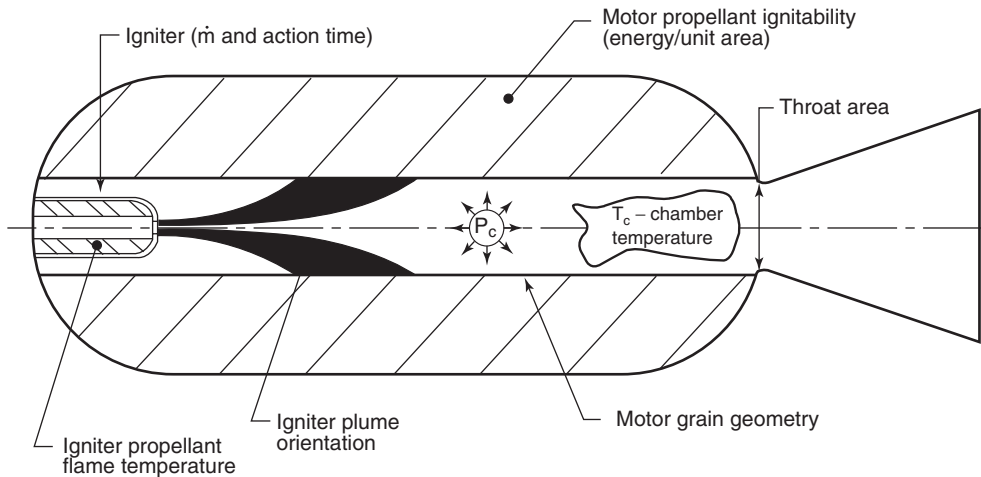


Figure 6. A typical solid propellant rocket motor with variables that affect ignition and grain heating.

There are four main components in an ignition system: the safety device, the electroexplosive device, the booster charge, and the igniter main propellant charge. The safety device provides the electrical and mechanical safety features needed to keep stray electrical current from causing inadvertent ignition. For planned ignition, electrical arming commands are sent to the safety device that mechanically and electrically arm or align it. In this armed condition, an electrical firing signal can then start a chain of events that culminates in motor ignition. The electrical firing signal is sent to an electroexplosive device called an initiator. The output from the initiator typically ignites a booster charge. The output from the booster charge then ignites the igniter main propellant charge that in turn ignites the solid propellant in the rocket motor.

Several methods are used for initial igniter sizing. One simple method uses an igniter coefficient. This is a ratio of the mass flow rate of the igniter gases to the throat area of the motor. Values of 0.25 lbm/sec per square inch of throat area are typical.

Future of Solid Propellant Rockets

Solid rocket motors have a long history of service because of their inherent characteristics: they are low cost, simple, reliable, compact, storable, and deliver very high thrust/weight. They have historically demonstrated the capability to support a wide range of propulsion applications. The entire U.S. ICBM and Submarine Launched Ballistic Missile fleet converted to solid propulsion in the 1950s and 1960s because of their high performance, high reliability, responsiveness, low cost, and packaging efficiency. Similarly, all land, air, and sea launched tactical missiles are solid propulsion based.

As science, engineering, and manufacturing have advanced, so have these rockets. Today, these rockets have very highly developed performance. However,

in the future, new structural materials will offer higher temperature capability and lighter weight. New propellants will offer more energy. New analytical techniques, combined with ongoing laboratory work continues to enhance our understanding of these motors. New tooling approaches for case, nozzle, and propellant grain will reduce cost and allow rapid design changes. Improved process controls will improve reliability, repeatability, and reduce cost. Improved case, nozzle, and insulating materials will increase reliability, reduce cost, and increase performance. New propellants will be more energetic, less sensitive, and less costly. Because of the basic strengths of solid propellants and because science and technology continues to build on these strengths, solid propellants can be expected to serve well into the future.

BIBLIOGRAPHY

1. Halliday, D., and R. Resnick. *Fundamentals of Physics*. Wiley, 1970.
2. Kruse, R.B. Fundamentals of Solid Propellant Rocket Motors. Paper at Thiokol Corporation Workshop, December 7–18, 1998.
3. Oberth, A.E. *Principles of Solid Propellant Development*, CPIA Publication #469. 1987.
4. Solid Rocket Motor Nozzles. NASA SP-8115, NASA Monograph, June 1975.
5. Wong, E.Y. Solid Rocket Nozzle Design Summary. AIAA Paper, #68-655, *AIAA 4th Propulsion Specialist Conf.*, June 10–14, 1968.
6. *Aerotherm Charring Material and Ablation Program* (CMA); Report No. UM-70; April 1970.
7. *Aerotherm Chemical Equilibrium Program* (ACE). Report No. UM-70-13, May 1970.
8. Canfield, A.R., E.E. Anderson, and G.E. Nichols. Space Shuttle Nozzle Development, AIAA Paper #78-951, *AIAA/SAE 14th Joint Propulsion Conf.*, July 25–27, 1978.
9. Solid Rocket Thrust Vector Control, NASA Space Vehicle Design Criteria Monograph, Chemical Propulsion, NASA SP-8114. December 1974.
10. Thrust Vector Control. Thiokol Corporation, AIAA Solid Rocket Technical Committee Lecture Series, *AIAA Aerosp. Sci. Meet.*, Reno, NV, January 9–12, 1995.

ROBERT R. BENNET
GEORGE A. BERKLEY
DOUGLASS B. COOK
BRIAN ALLEN
ROBERT P. GRAHAM
ARLEIGH P. NEUNZERT
RONALD W. LYMAN
BILLY H. PRESCOTT
C. PAUL PULVER
STEVEN R. WASSOM
NICHOLAS J. WHITEHEAD

JOSEPH J. KLIGER
DONALD R. SAUVAGEAU
ROBERT L. CRIPPEN EDS.
Thiokol Propulsion, Inc.
Brigham City, Utah

SPACE LIFE SCIENCES

Definition

The term space life sciences refers to the set of sciences that study the biology of living things under conditions of spaceflight. The synonymous term, space biology, coined in 1920 by K. Tsiolkovsky, is more common in Russia (USSR). The origin and development of space life sciences as a discipline are the result of certain advancements in modern science and technology and are associated with the development of jet propelled flight vehicles that can overcome the force of gravity to ascend beyond the bounds of Earth's atmosphere.

Space biology arose at the intersection of many scientific disciplines: various branches of biology and medicine, physics, chemistry, astronomy, astrophysics, geology, geophysics, geochemistry, radiology, and mathematics. It is indissolubly bound up with cosmonautics and space medicine.

Space medicine is the area of general medicine that studies what happens to human physiological processes under the unique conditions of space flight to develop means and methods for maintaining the health and performance capacity of crews of spacecraft and space station pre-, in- and postflight. As one of the branches of occupational medicine, space medicine attempts to provide the answers to a number of questions confronting humanity as a result of scientific and technological progress. These answers are derived from a very large medical knowledge base relating to various areas of theoretical and clinical medicine such as space physiology and psychophysiology, space hygiene, space radiobiology, and medical expertise.

In the USSR/Russia, space medicine has been seen historically as a composite discipline, encompassing space physiology, which addresses scientific issues involving all of the factors that affect humans in space, and space medicine proper, which addresses the practical issues involved in maintaining cosmonaut health and performance capacity. Thus, in many references in the literature, the term space medicine is used in its broad sense to refer to the combination of space medicine narrowly defined and space physiology.

The mission of space physiology is to find out how much it is possible to decrease comfort and stress human physiological functions, still induce only physiological changes that are merely adaptive rather than pathological and leave performance capacity and job efficiency undiminished. In the light of the high cost of flight time in space, the constant time deficit and the serious potential consequences of human error, it is critical that cosmonauts' performance capacity be maintained at a high level.

Goals and Objectives

The goals and objectives of the principal space life sciences are presented in Table 1.

Meeting the challenges confronting space life sciences involves efforts in the following major research directions:

- investigation of the effects of weightlessness and other spaceflight factors, including the combined effects on physiology as a whole and on fundamental

Table 1. Objectives and Goals of Space Life Sciences

<i>Space biology</i>	
To conduct fundamental biological research devoted to comprehensive study of space and celestial bodies as a unique living environment and of the effects of factors of the space environment, including spaceflight factors, on various forms of life	<ol style="list-style-type: none">1. Study of the biological effects of various spaceflight factors on terrestrial life forms2. Investigation of the possibilities for life to exist under extraterrestrial conditions, including other planets (exobiology)3. Study of the mechanisms governing changes in the physiological functioning of humans and animals under exposure to spaceflight factors (space physiology)4. Development of life support systems (LSS) for use by terrestrial life forms under spaceflight conditions
<i>Space medicine</i>	
To support maximum possible physical and psychological well-being, high performance capacity, job efficiency, and increased reliability of human performance under the unique conditions of spaceflight and after its completion	<ol style="list-style-type: none">1. Creation of the requisite conditions to optimize human vital functions during space flight2. Monitoring, predicting, and correcting physiological changes that occur as a result of the impossibility of fully or partially replicating terrestrial conditions3. Maintaining human health during all phases of preparation for flight, flight itself, and after flight completion (selection, observation, diagnosis, treatment, and prevention of pathologies)4. Ensuring the safety and increasing the efficiency of spaceflights; providing medical and ergonomic measures to increase the reliability of human-machine interactions.

processes at the level of individual organs, systems, tissues, cells, and at the subcellular level;

- study of the mechanisms by which living things, including human beings, adapt to exposure to spaceflight factors in experiments aboard spacecraft and in ground-based simulations;
- investigation of individual differences in regulating the psychophysiological status of cosmonauts during exposure to the extreme environment of space;
- work on the fundamental problems of space physiology, gravitational biology, ecology, and exobiology.

Major Phases in the Evolution of Space life Sciences

Development of the Underlying Knowledge Base and Formulation of the Issues Involved in Preparing for Spaceflight. The scientific goal of preparing for and actually accomplishing a spaceflight was formulated on the basis of an existing scientific knowledge base and previously developed methodological

approaches. The knowledge needed to formulate the goal of preparing for spaceflight came from the following sources:

- expeditions to high mountain regions;
- underwater immersion;
- physiological experiments (using centrifuges, barochambers, etc.);
- balloon and, later, aircraft flights.

The sources of the research results that form the knowledge base for space medicine go as far back as the Renaissance. Galileo first described the effects of gravity on living things in 1638. Newton made the first proposal to use rocket-powered vehicles for scientific space travel and suggested methods by which man-made satellites could be inserted into orbit around Earth in 1687. Boyle used a barochamber to conduct experimental investigations of the effects of diminished atmospheric pressure on animal physiology from 1660–1692. A great deal of knowledge was generated by research conducted, without specific consideration of spaceflight, in the eighteenth through twentieth centuries in areas such as pathological physiology, high-altitude physiology, underwater and diving medicine, and aeronautics, and aviation medicine.

Of all of the preexisting biomedical disciplines, aviation medicine had the most important role in making piloted spaceflight possible. At the intersection of medicine and technology, this discipline generated information about the physiological effects of high levels of atmospheric decompression and dynamic flight factors and acquired technical expertise in developing means of preventing the adverse effect of these factors. All of these achievements were subsequently used in preparing for the first spaceflight.

Study of the laws governing physiological reactions to extreme environmental conditions enabled aviation medicine to develop a methodology for analyzing physiological reactivity to the effects of an extreme factor and for preventing this factor from having adverse physiological effects. This methodology was based on the goal of supporting optimal human performance capacity in extreme environmental conditions. Use of this methodology gave researchers a tool that enabled them to solve the analogous problem with regard to spaceflight conditions. This achievement of aviation medicine was one of the most important contributions to the knowledge base for space medicine.

Piloted flights into the stratosphere (1933–1934) in the pressurized gondola of the USSR and Osoaviakhim high-altitude balloons (reaching an altitude of 18,600 m and 22,000 m, respectively) functioned as a prelude to future piloted flights on spacecraft. These stratospheric balloon flights demonstrated empirically that the problem of operational medical flight support was indeed soluble. Preparing for these flights allowed experts to gain a great deal of practical experience in creating life support systems (LSS) for conditions equivalent to those in space. Stratospheric flights provided the impetus for developing pressure suits and pressurized aircraft cabins for altitudes above 10,000 meters. Moreover, they generated the first information on the biological effects of a previously unstudied factor of the space environment—primary cosmic ionizing radiation. The papers delivered in March 1934 at an All-Union Conference on the Study of the

Stratosphere held in Leningrad formulated a number of biomedical problems related to the study of spaceflight. The goal of human penetration into the upper atmosphere was scientifically formulated for the first time and publicly supported at this conference. Consequently, a scientific center was set up under the aegis of the USSR Academy of Sciences to achieve this goal.

Phase One: Creation of an Independent Scientific and Applied Discipline.

The technical capacities needed to initiate practical research to prepare for piloted spaceflight began to be developed in the late 1940s. During this period, ground-based experiments; experiments conducted on aircraft, rockets and balloon flights; and the theoretical analysis that preceded them were concentrated in three areas: study of the effects of gravity within the broad range of 0 g to several g's, study of the biological aspects of the effects of cosmic radiation, and development of a LSS for pressurized cabins. Solving the initial applied problems involved in preparing for piloted spaceflights required efforts to generate absolutely new information on the physiological effects of space environment factors and also a great deal of work to analyze already existing databases in related disciplines. In addition, to guarantee the maximum possible safety of piloted spaceflight, simulations and field experiments had to be conducted to verify empirically whether it was appropriate to extrapolate already established laws to conditions in space.

In 1949, under the direction of A. Blagonravov, the Soviet Union began to implement a program for using the R-1 (V-1) rocket to study physical processes in near-Earth space. In 1950, a group headed by V. Yazdovsky was formed within the Institute of Aviation Medicine to work on the biomedical aspects of high-altitude rocket flights. The first series of experimental tests (1951) involved six launches (of which two failed) of the R-1C rocket that carried dogs to an altitude of approximately 100 km. The first geophysical rocket, R-1C, carrying the dogs Dezik and Tsygan, was launched on 22 July 1951 from the Kapustin Yar Cosmodrome, ascended to an altitude of 100.8 km, and landed successfully. This was the world's first successful flight of an animal on a rocket. The results of this effort included selection of biological subjects for spaceflight research, development of an animal pressurized cabin LSS, development of sensors and monitoring devices and of methods for studying animals' physiological functions that could be used under conditions of rocket flight, and study of the nature and severity of the effects of high atmospheric flight on the physiological functioning and behavior of animals.

The second set of experimental trials (1954–1956) involved launching nine R-1E and R-1F rockets carrying a total of 12 dogs to altitudes up to 110 km. This research showed that animals satisfactorily tolerate conditions during powered flight and the subsequent 3–6 minute period of weightlessness and that the LSS and emergency catapult system (for use from a variety of altitudes), as well as the methods for recording physiological functions and the filming system that had been developed, performed adequately.

The next major step was the launch of the second nonrecoverable Earth satellite vehicle (ESV-2) carrying the dog Layka on 3 November 1957. The first orbital flight of a living creature made it possible to test the LSS, the methods that had been used to select and prepare flight animals, and methods for studying a number of physiological functions in flight and for transmitting biomedical information from the flight vehicle to Earth. Biomedical studies conducted on

ESV-2 supplemented material obtained from vertical rocket launches and provided researchers with the essential experimental data to confirm initial hypotheses that exposure of living things, including humans, to space does not cause them any harm. This experiment demonstrated that it is possible, in principle, for a higher animal to complete a spaceflight in near-Earth space.

The results obtained from studies of a broad range of biological subjects on flights of rockets and the first Earth satellite vehicles were meant to be used to prepare for piloted spaceflight (1950–the early 1960s) and indeed demonstrated that there are no biological limits on living under flight conditions in near-Earth orbit.

Getting ready for the first piloted spaceflight entailed solving a number of new problems and developing a number of new research directions. First of all, the physical properties of near-Earth space had to be studied, and a number of issues associated with biomedical, ballistic, and navigational support of piloted spaceflights had to be resolved. It was essential to solve the problem of crew life support (air regeneration, water supply, and maintenance of normal temperature conditions in the crew cabin), ensure their safe return to Earth from orbit, maximize the likelihood of landing in the predetermined region, limit acceleration during powered flight, and provide reliable thermal insulation for the spacecraft's reentry into the dense layers of the atmosphere.

In the Soviet Union, this research was conducted under the direction of V. Yazdovsky and O. Gzenko during animal flights on the R-2A (five flights) and R-5A (nine flights) geophysical rockets to altitudes from 200 up to 473 km (1957–1960). A total of 14 dogs flew on these missions, some of them two to four times each, as did other animals—rabbits, rats, and mice. These experiments showed that animals do not display any notable disruption of physiological status or behavior in the weightlessness induced by geophysical rocket flights.

The successful advance of research in the Soviet Union, as well as the development of a reliable launch vehicle—the R-7—in 1958–1959, made it possible to move on to the final phase of preparation for the first piloted spaceflight, the development of the Vostok piloted spacecraft, and implementation of a cosmonaut selection and training program. These efforts were conducted at virtually the same time and in parallel.

The previous rocket experiments that established that higher animals can well tolerate short-term exposure to spaceflight factors were merely the starting point for solving a wide range of problems pertaining to selection. This problem was so critical because the overall success and results of spaceflights, in many respects, would depend on cosmonaut physiological status. Incorrect evaluation of an individual's functional capacities or an undiagnosed disease could have posed a threat to the cosmonaut's life or resulted in the flight mission's failure. It was very difficult to conduct an adequate evaluation, primarily because it was impossible to replicate the entire set of factors to which a human being is exposed during spaceflight. Development of the cosmonaut medical selection system used the many years of experience accumulated by aviation medicine in qualifying flight crews for various types of aircraft. Naturally, this system could be used for selecting cosmonauts only if the differences in the professional tasks performed by cosmonauts and the characteristics of their environment were taken into account.

In-depth medical examinations and special tests involving exposure to parabolic aircraft flights, centrifuges, barochambers, thermal chambers, rotating tilt tables, and anechoic chambers resulted, in December 1959, in the selection of the 20 members of the first USSR cosmonaut team. During flight training, a great deal of emphasis was placed on biomedical preparation, including conditioning of the candidates' muscle, cardiovascular, and vestibular systems. The principal goal was to prepare the cosmonauts physiologically for the combined effects of such unaccustomed factors as acceleration, weightlessness, vibration, motion sickness, and nervous stress. Thermobarochambers, centrifuges, anechoic chambers, training simulators, and vibration benches were used extensively in this conditioning program.

In April 1960, the Soviet Union began a series of (unpiloted) flight tests of the Vostok piloted spacecraft, called a piloted orbital spacecraft (POS). In 1960–1961, there were a total of seven orbital spacecraft launches, of which only three were complete successes. On 19 August 1960 the third flight of the Vostok orbital spacecraft (receiving the mission designation POS-2) was successfully completed. For the first time ever, the dogs Belka and Strelka and other biological subjects completed a 1-day, 17-pass orbital flight and returned safely to Earth. After this flight, specialists could, for the first time, study the effects of spaceflight conditions on living creatures and conduct meticulous physiological, genetic, and cytological analyses of the animals and other biological subjects after they had been in space for 25 hours. Two more launches of the orbital spacecraft (POS-4 carrying the dog Chernuska and POS-5 carrying Zvezdochka) immediately before the first piloted flight, confirmed that it was feasible for a human being to fly in space. Investigations conducted on these launches showed that the danger from meteors in near-Earth space on the intended flight trajectory was virtually nil and that the level of radiation beyond the radiation belts was not as high as it is during solar flares.

The first piloted space flight took place on 12 April 1961. Yuriy Gagarin completed one pass in near-Earth orbit for a flight of 1.8 hours (108 minutes). Throughout the entire flight, his physiological parameters remained within the limits that could be described as the zone of normal reactions to unusual environmental factors. The results of this flight demonstrated that it is possible, in principle, for human beings to fly in space and that humans can adapt to the unaccustomed state of weightlessness without losing their performance capacity and ability to orient themselves spatially. A meticulous analysis of scientific materials obtained during the flight enabled comprehensive tracking of the cosmonaut's physiological functions over time and the parallel characteristics of the spacecraft systems' operations. This made it possible to improve the spacecraft equipment and make the necessary adjustments in the cosmonaut training program. The main result was that the flight demonstrated that, in principle, there are no biological constraints on flights in near-Earth space. This opened the door to a gradual increase in the duration of the human presence in space and in the amount of work performed therein, and thus completed the process of the initial development of space medicine as a science and as an independent branch of medical practice.

Phase Two: Research on Short-Term Piloted Flights in Near-Earth Space.

Yuriy Gagarin's successful flight was followed by a large number of flights by

Soviet cosmonauts and U.S. astronauts. In August 1961, G. Titov spent more than a day in space on board Vostok-2. The cosmonaut worked, slept, and ate and thus demonstrated that the basic diurnal rhythms of life are maintained in space. During his flight, he made meteorological and geophysical observations, took the first movie from space, and controlled the spacecraft. Extremely valuable scientific material was generated, making it possible to draw conclusions about the effects of spaceflight factors on human physiology during an entire diurnal cycle. Titov also demonstrated that it is possible to perform mental (operator and research) work in space. This was the first flight on which a cosmonaut complained of malaise. G. Titov noted symptoms of motion sickness (dizziness, nausea) when he moved his head, which were more severe during the second pass. The first piloted space flights demonstrated that human beings can live and work in space. However, the issues of how long and how well they could do this required further biomedical research in space.

On these premises, exploration began of the more complex systems and methods required for systematic scientific research on human activities in space. In particular, the following problems were studied: can a human being function in space during a prolonged period; can women tolerate space flight; can multiman crews operate in space; can spacecraft rendezvous and dock; is work in open space possible.

In August 1962, the Soviet Union conducted the first multiman space flight (A. Nikolayev, P. Popovich); in June 1963, the first female cosmonaut, V. Tereshkova, spent 3 days in space. The multiman Voskhod spacecraft flew for the first time in October 1964, carrying a crew of three men, including the first physician-cosmonaut, B. Egorov. The first extravehicular activity (EVA) (A. Leonov, 1965) took place during the flight of Voskhod-2.

The next important step was the USSR's creation of the Soyuz spacecraft in the mid-1960s. Soyuz was intended for further development of processes of autonomous navigation, performance of transport operations (delivery of cosmonauts to space station), conduct of scientific-technical and biomedical experiments in near-Earth space, and for performance of multilevel scientific and applied tasks. The new spacecraft differed from Vostok and Voskhod by virtue of its ability to perform orbital maneuvers and to approach and dock with other spacecraft. The Soyuz cabin was divided into two modules: the reentry module (for returning cosmonauts to Earth) and the orbital module (for conducting scientific research and cosmonaut daily activities and sleep).

The biomedical studies conducted in the 1960s during short-term piloted spaceflights had demonstrated that humans could safely spend 2 or 3 weeks under conditions of weightlessness and perform EVA. After the cosmonauts returned to Earth, certain changes were observed, which, it seemed, increased in severity with increasing duration of space missions. This led to the development of methods to prevent the adverse physiological effects of weightlessness; however, these measures were not actively adopted in space medicine until 1970. The 18-day flight of Soyuz-9 (A. Nikolayev, V. Sevastyanov, 1970) was very significant for the solution of a number of biomedical problems. This flight generated data on the reactions of cardiovascular and musculoskeletal systems to prolonged exposure to weightlessness, studied the role of physical exercise, and investigated the characteristics of the postflight recovery period.

The cosmonauts had problems tolerating what at that point was the longest ever exposure to weightlessness. After the flight, it was found that they exhibited significant atrophic changes in their muscles and orthostatic intolerance, which required medical rehabilitation for a number of days. The experience gained during this flight provided valuable material for studying whether it would be possible for humans to spend longer periods in space. The significant physiological changes noted in the cosmonauts stimulated work to develop systems of targeted preventive measures and enhanced medical examinations and their adoption in operational medical flight support to ensure reliable and safe piloted spaceflights of increasing duration.

Phase Three: Research on Long-Term Piloted Spaceflights in Near-Earth Orbit. The results of short-term piloted flights during the late 1960s suggested that cosmonauts did not undergo any physiological changes that went beyond the bounds of nonspecific responses to extreme environmental conditions or that would be an impediment to further increases in flight duration. This conclusion instilled optimism regarding the planning of future long-term spaceflights and was an impetus to development of a spacecraft of a new type—the piloted orbital space station. The relatively large size of these stations, which could grow by docking with additional modules, made it possible for them to carry equipment for extensive biomedical investigations and for preventing the adverse effects of spaceflight factors. Moreover, these space stations provided comfortable conditions for living and personal hygiene and diminished the physiological effects of severe constraints on motor activity characteristic of smaller spaceflight vehicles.

The strategy developed by Soviet experts involved a sequential, gradual increase in the duration of human exposure to space without damage to health while maintaining satisfactory capacity for performing one or another flight mission. The tactics for implementing this strategy were determined by the results of studies conducted on the flights of piloted orbital spacecraft and unpiloted biosatellites, ground-based experiments, and data from general physiology and the practice of medicine. Continuous accumulation of knowledge on the physiological effects of spaceflight factors made it possible to improve the system for selecting and training cosmonauts, the measures used for medical monitoring and predicting of health status and for preventing the adverse effects of weightlessness, and rehabilitative measures to be used after flights of long duration.

During this phase, each successive piloted spaceflight represented a mean increase in duration of 30–40 days over the previous one. This pattern was dictated by the reliability of medical prediction based on the results of all previous flights. The psychological factor was also considered of major importance in planning increases in piloted flight duration. Thus, increasingly long flights of 96, 140, 175, 185, 211, and 237 days were implemented successively. The work of cosmonaut physician O. Atkov on the 237-day flight in 1984 was of great significance. The presence of a physician on the station made it possible to validate new modes and schedules of physical exercise and to conduct comprehensive studies of the cardiovascular system, which included echographic studies.

The 1-year flight on the Mir Space Station (V. Titov, M. Manarov, 1987–1988) was a significant milestone. This accomplishment was made possible by implementing an extensive set of planned investigations and successfully solving

a number of challenges involved in medical support of long-term flights. New exercise schedules on the bicycle ergometer and treadmill—the major prophylactic countermeasures—were validated, as were new schedules for provocative tests using lower body negative pressure. A whole series of neurophysiological, hematological, and biochemical studies performed by cosmonaut physician, V.V. Polyakov, during the last phase of this flight made it possible to expand significantly the potential for diagnosing and correcting functional disorders arising in cosmonauts during flight.

The culmination of long-term flights on space station Mir was the record-setting super-long-term, 438-day flight of cosmonaut-physician V. Polyakov (1994–1995). The major result of this flight was the demonstration that cosmonauts can retain their health and performance capacity on a flight comparable in duration to a Mars mission, and that the high functional physiological capacities of the crews that worked with him were also retained.

A new chapter was opened, when the International Space Station (ISS) developed through the joint efforts of Russia, the United States, the European Union, Japan, and Canada, went into operation in Earth orbit. Considering the growing demands made on operational medical support of upcoming piloted space flights (see Biomedical Support of Piloted Spaceflight), the following major directions of research will be performed on the ISS:

1. evaluation of the physiological effects of spaceflight factors as a function of their combination, intensity, and duration and the study of individual adaptive and re-adaptive physiological reactions;
2. study of the mechanism through which the structure, functions, and behavior of living things change, whether or not they are reversible, and their remote consequences;
3. identification of the most information-rich physiological criteria for ongoing and predictive evaluation of cosmonaut status, pre-, in- and postflight, for use, among other purposes, to optimize prescribed dosages of prophylactic interventions;
4. development of a physiological rationale for techniques to increase specific and nonspecific physiological resistance to a particular combination of spaceflight factors and working and living conditions.

Phase Four: Preparation for Interplanetary Spaceflights. At present, space medicine is at the threshold of the next stage in its evolution, associated with supporting autonomous work by cosmonauts, for example, at a scientific lunar base or on board an interplanetary spacecraft. The novelty of the challenges presented by such flights results from the significantly greater crew autonomy arising, at least in part, from the impossibility of quick or premature return to Earth in cases of emergency or illness and the need to increase significantly the reliability of technical and medical systems. Russia, the United States, and a number of other countries are theoretically analyzing the challenges facing them and are performing experimental work on particular biomedical aspects of autonomous flight. One of the initial phases here involved long-term flights, including the 438-day flight on Mir, which demonstrated that,

in principle, there are no biomedical considerations that would prevent a Mars mission. It is assumed that the ISS shall also become a test bed for developing various aspects of interplanetary flights.

The specific aspects of interplanetary spaceflights that require new approaches to the organization of medical support systems are enumerated in Table 2, using a Mars mission as an example.

A piloted flight to Mars demands solving a number of physiological problems resulting from long-term exposure to weightlessness. The symptoms and mechanisms of the physiological changes occurring under these conditions in prolonged flights have been studied in relatively great detail; however, the prophylactic measures that have been developed require further improvement. In particular, the possibility of using artificial gravity (AG) for this purpose will be investigated, if nongravitational prophylactic measures do not prove effective enough. At the same time, the use of AG could lead to the occurrence of a number of physiological problems from exposure to a rotating system: development of sensory conflicts, difficulty in motor orientation, or adverse effects on the vestibular system. The problem of the combined effects of factors involved in a flight to Mars remains very critical.

On the highly autonomous spaceflights of the future, for example, flights of Mars and other planets of the Solar System, cosmonaut meals will no longer consist solely of stores of food brought from Earth. Instead, the crew's nutritional system will have to be based on food substances and products produced on board, including those based on nontraditional food sources.

The procedures for crew interactions with ground control services will also have to be revised completely. Because of the delay in radio signals on the Earth–Mars path, it will not be possible for ground-based flight control services to react immediately to events on board, and this will diminish their function mainly to one of consulting and support; the entire responsibility for making moment-to-moment decisions will be borne by the Mars spacecraft crew. Thus, the risk of the Mars mission is significantly higher than that for cosmonauts in near-Earth

Table 2. **Specific Features of a Piloted Flight to Mars**

-
- Long-term (no less than 2 years) crew residence in an artificial living environment, which would lead to biological and chemical microcontaminants accumulating in the atmosphere, the formation of an unusual microbial community inside the spacecraft, and the possible deviation of microclimatic parameters from those considered safe
 - Possible exposure to galactic cosmic radiation without screening by Earth's magnetosphere
 - Long-term continuous exposure to a hypomagnetic environment and solar ultraviolet radiation
 - Diminished possibility for ground control services to respond quickly to what is happening on board as a result of 15–30 minute signal delays
 - Impossibility of emergency or premature crew return to Earth or replacement of an ailing crew member
 - Cosmonaut exposure to hypergravity during landing, stay on Mars, takeoff from Mars, and landing on Earth
 - Need for the crew to live and work together for a long period in isolation, possibly leading to development of psychological incompatibility and psycho emotional stress
-

orbit. Thus, implementation of these missions will have to be preceded by intensive and in-depth investigations in space physiology, psychology, and radiobiology as well as by the development of LSS equipment and other technical spacecraft equipment with significantly higher reliability and maintainability than those in use today.

Major Research Directions

The scientific data needed to solve the problems confronting space medicine have been generated by biomedical experiments conducted on space flights, ground-based simulation studies, and practical experience with operational medical flight support, as well as by advances in general physiology and medicine.

Experiments on Board Piloted Spacecraft. The 40-year period during which piloted spacecraft have been flying has made it possible to accumulate unique experience in solving biomedical problems to ensure the safety and efficacy of spaceflights of ever increasing duration. During this period, the duration of piloted space flights has increased to 12–14.5 months for men and up to 6 months for women. The successful implementation of biomedical research programs in space has significantly facilitated this progress. Among those who have spent the most time working in orbit, we should mention S. Avdeyev (750 days on three spaceflights), V. Polyakov (679 days on two spaceflights), A. Soloviev (653 days on five spaceflights), V. Afanasyev (547 days on three spaceflights), and A. Viktorenko (489 days on four spaceflights).

An extensive program of biomedical research and experiments has been implemented on the Salyut-3, -4, and -5 orbital stations that were inserted into orbit between 1974 and 1976. The total duration of work performed by six cosmonaut crews on these stations was 176 days. Much emphasis was placed on developing the optimum physical exercise programs using exercise machines and on improving the work-rest schedule to increase the postflight resistance of the cardiovascular system to the effects of Earth's gravity.

The flight of Salyut-4 witnessed the first use of the "Chibis" vacuum suit, which applies lower body negative pressure to simulate the effects of Earth's gravity on the blood system and thus to counteract excess blood flow to the head. The station was also fitted with a rotating chair for studying vestibular function and a bicycle ergometer (in addition to the previously used treadmill). Another innovation was the "Tonus" apparatus for electric stimulation of separate groups of muscles. Additionally, cosmonauts on Soyuz-18 were fed a diet containing increased levels of salt and additional liquid. This regimen proved very effective in increasing their tolerance for conditions on return to Earth.

The Salyut-6 space station was inserted into orbit in September 1977 and remained in orbit for more than 4.5 years. During this time, it was inhabited by 29 cosmonauts comprising 16 crews (five prime crews and 11 visiting crews), who conducted a large number of scientific and technical investigations, including more than 1600 biomedical and biological studies.

The last station in this series—Salyut-7—was inserted into orbit in April 1982 and remained there for more than 4 years. Ten crews worked on Salyut 7, including five prime crews and five visiting crews, a total of 23 cosmonauts,

including the second Soviet cosmonaut physician, O. Atkov, who conducted a large number of ultrasound studies of cosmonauts' cardiovascular systems, investigated mineral and carbohydrate metabolism, and determined the optimal parameters for physical exercise and loading.

In February 1986, the Soviet Union launched the core module of the third-generation space station, Mir, which formed the basis for a multimodule orbital complex. A total of 44 crews (28 prime and 16 visiting crews), comprising a total of 104 cosmonauts, including 63 foreigners, worked on board this station. A series of international projects were implemented on board with the help of citizens of the United States, France, Germany, the United Kingdom, Austria, Syria, Bulgaria, Slovakia, Afghanistan, and representatives from the European Space Agency (ESA). In addition, nine crews visited Mir on the U.S. Space Shuttle, including citizens of the United States, Canada, France, and representatives from the ESA (a total of 50 people). Mir was the first international piloted space station.

A permanent and important component of Mir flight programs was the broad spectrum of biomedical research studies (a total of 1759) that generated new information on the mechanisms underlying the changes that occur in various human and animal functional systems under conditions of weightlessness and on the characteristics of the processes of physiological adaptation to space flight and of readaptation to conditions on Earth. Mir experiments demonstrated that an entire life cycle could take place in weightlessness. Progress was made in studying the habitability of piloted spacecraft complexes (sanitary/hygienic, microbial, and radiation physics studies); experience was accumulated in the sanitary and hygienic support of crews during long-term habitation of space stations and during contingency modes of LSS operation.

During the 438-day flight alone, V. Polyakov performed approximately 1000 biomedical analyses and tests in the following areas:

1. use of clinical physiological and laboratory analyses to investigate the mechanisms underlying adaptation of human functional physiological systems to conditions on long-term flights;
2. study of the problems of habitability on long-term space flights and optimization of crew living conditions on board;
3. improvement of medical systems ensuring cosmonaut safety on spaceflights.

The long period during which Mir was used (15 years) made it possible to develop unique expertise in solving biomedical problems to support the safety and efficacy of spaceflights of increasing duration. In particular, the principles governing the evolution of microflora during multiyear use of inhabited facilities with an artificial environment were established, medical and technological risks were defined and classified, means and methods for monitoring and supporting ecological crew safety and for protecting the spacecraft interior and equipment from biodegradation were developed, and other scientific and hygienic problems were solved.

An extensive and multifaceted program of fundamental biological research on plants, birds, and amphibians was conducted on the Salyut and Mir stations.

This program included studying the growth and development of a variety of life forms to develop biological life support systems for future flights. Thus, in the *Chlorella* experimental program on board Salyut-6, it was first demonstrated that weightlessness had no primary biological effect on an actively growing culture of one-celled algae, either at the level of the organism, or at the level of interactions within the “organism–environment” system. Multiyear studies conducted on Mir (the *Oranzherya* and *Inkubator* experiments) first demonstrated that living organisms could undergo a full developmental cycle under conditions of weightlessness; organogenesis was fully demonstrated in mammals and birds, and the full developmental cycle of wheat from seed to mature plants was realized.

Experiments with a model microecosystem (the *Aquarium* and *Aquarium-M* experiments) on Mir and on the Bion biosatellites first demonstrated that weightlessness does not affect the functioning of an “algae–bacteria–fish” microecosystem as a whole. All changes in weightlessness that occurred in the system were similar to those it underwent under normal gravity. The growth, development, and reproduction of populations of one-celled algae within the system proceeded normally in space. Spaceflight factors failed to impact algae productivity or their functioning as the autotrophic component of the microecosystem. These results are of important fundamental and applied significance for the development of scientific methodological principles to underlie the design and subsequent implementation of hybrid biological–physical–chemical life support systems (controlled ecological LSS).

As a result of the research conducted on space stations, the major risk factors for piloted spaceflights were defined, and the specific and nonspecific laws of human adaptation to space flight factors were studied. The major physiological systems most subject to changes under conditions of long-term flight were identified, and effective means and methods for preventing undesirable physiological changes in response to spaceflight factors were developed. This research made a significant contribution to solving problems in gravitational biology—the science that studies the effects of gravity on life forms, including human beings. All of this activity culminated in the implementation of comprehensive clinical–physiological studies of the functions, regulation, and structure of various systems pre-, in- and postflight. The results generated are of great importance for the practice of space medicine and also for solving fundamental problems in gravitational physiology and life science as a whole.

Bion Program Research. Between 1973 and 1977, systematic biological and physiological investigations were undertaken as part of the Bion program on flights of 11 Kosmos series biosatellites to deepen understanding of the effects of spaceflight factors, especially weightlessness, on vital processes. These satellites were equipped to conduct flight experiments on various species of animals and plants. All of this research was directed by the Institute of Biomedical Problems and involved extensive domestic and international cooperation. A broad range of scientific disciplines, such as gravitational biology, physiology, developmental biology, cellular and radiation genetics, metabolism, morphology, histology, hematology, and immunology, were involved. Scientists from the United States, France, Canada, Holland, Bulgaria, Hungary, Germany, Poland, Rumania, Czechoslovakia, and China participated in this joint research.

The experimental subjects the biosatellites carried included cell and tissue cultures, one-celled organisms, plants, insects, amphibians, fish, reptiles, bird eggs, and mammals—a total of approximately 40 species of living things at different levels of evolution and ontogenesis. A list of Bion flights is provided in Table 3.

The use of a large number of highly varied biological subjects for research made it possible to obtain statistically significant data on the direct and mediated physiological effects of weightlessness per se (i.e., without social and psychological overlays and without the modulating effects of the prophylactic countermeasures employed on every piloted flight). The investigations performed revealed the universal significance of the gravitational factor in the formation of the structure and functions of living systems and laid the foundations for a new scientific discipline—gravitational biology. It was demonstrated that weightlessness is the major etiological factor that induces physiological changes on spaceflights in near-Earth orbit. The experiments studied the mechanisms underlying the biological effects of weightlessness on the cellular, systemic, and organismic levels, identified the principles governing adaptation of living systems to weightlessness and readaptation to normal gravity, evaluated the effects of artificial gravity, and investigated the biological effects of the heavy components of galactic radiation and the combined effects

Table 3. **Flights in the Bion Program**

Biosatellite	Year	Duration (days)	Biological subjects
Kosmos-605	1973	22	45 male rats, tortoises, fruit flies, meal worms, crown galls of carrot plants
Kosmos-690	1974	20.5	35 male rats, tortoises, fruit flies, pine seeds, fungi, bacteria
Kosmos-782	1975	21	45 male rats, fruit flies, fish roe, yeast, carrot crown galls
Kosmos-936	1977	18.5	30 rats (male and female), fruit flies, higher and lower plants
Kosmos-1129	1979	18.5	37 male rats, eggs of the Japanese quail, higher and lower plants, cultures of mammalian cells, carrot crown galls
Kosmos-1514	1983	5	2 monkeys, 10 male rats, guppies, crocus, corn seed sprouts
Kosmos-1667	1985	7	2 monkeys, 10 male rats, fruit flies, guppies, newts, higher plants
Kosmos-1887	1987	14	2 monkeys, 10 male rats, insects, guppies, newts, planaria, higher plants
Kosmos-2044	1989	14	2 monkeys, 10 male rats, guppies, crocus, corn seed sprouts
Kosmos-2229	1992–93	12.5	2 monkeys, frogs, newts, insects, cell and tissue cultures, planaria, seeds, plant sprouts
Bion-11	1996–97	12	2 monkeys, newts, crustaceans, insects, French snails, one-celled animals, seed sprouts

of cosmic radiation and weightlessness. The flight of biosatellite Kosmos-936 produced data for the first time showing that artificial gravity created by an onboard centrifuge can prevent the adverse effects of weightlessness. This made it possible to consider artificial gravity as a promising method for maintaining the optimal physiological status of human participants in long-term space-flights.

Ground-Based Simulation Experiments. It is well known that the reverse side of scientific and technological progress is an increase in the extreme factors to which humans are exposed. Thus, it is natural that various different scientific experiments have had to be performed to assess the possible adverse consequences of such exposure to enable the development of the necessary preventive or protective countermeasures. Possibly, this need has been especially critical in aviation, space, and underwater medicine, which must confront the most complex and extreme external conditions that can have a significant effect on the health and occupational performance of humans.

To resolve these issues, of course, experts have used the experience accumulated in many areas of medicine and the results of initial studies on laboratory animals. However, in many instances, experiments or tests have had to be performed on humans if problems are to be solved once and for all. In 40 years, several hundred such investigations were undertaken in ground-based laboratories creating simulations of the effects of various spaceflight factors (isolation in a pressurized chamber, hypokinesia with head-down tilt, and water immersion) on living things. These tests have lasted from several days to a year. The centrifuge made it possible to study the effect of acceleration across a broad range of values, the barochamber—the significance and role of the barometric factor and altered atmospheric gas composition; and apparatus producing various types of ionizing radiation—the effects of the radiation factor. Water immersion made it possible to replicate certain biological effects of weightlessness.

A medical-technical experiment of a year's duration involving three test subjects, which was conducted in the Soviet Union in 1967–1968, was important for the development of advanced LSS. This experiment investigated the possibility of long-term (up to 1 year) retention of normal human performance capacity under conditions of isolation in a pressurized chamber of limited size using water and oxygen regenerated from wastes and virtually totally dehydrated food. It addressed the characteristics of human interactions with the environment under these conditions, methods of medical monitoring, technological methods for designing the various modules, and other issues. During the experiment, the test subjects lived in an isolation chamber consisting of a living module and an experimental greenhouse connected to each other. This test of a closed-cycle LSS demonstrated that it is possible to live and work for a long period of time within such systems.

A 182-day experiment involving hypokinesia that was conducted at the Institute of Biomedical Problems in 1976–1977 in many respects was a prologue to the phase of long-term flights. The research methods and prophylactic countermeasures used in this experiment generated data, which, along with results of examinations of cosmonauts who at that time had participated in flights of moderate duration, allowed medical personnel to predict that it would be possible to increase flight duration further systematically.

The results of another experiment conducted in the same Institute in 1986–1987 made a significant contribution to the realization of the 327-day flight and the succeeding 1-year flight. This unique experiment involved long-term (370 day) hypokinesia for nine volunteer subjects. In many respects, this experiment made it possible to deepen and expand our understanding of the mechanisms by which weightlessness affects human physiology. Moreover, the procedure, schedules of physical exercise, and certain drugs whose efficacy it validated were subsequently used by cosmonauts on long-term flights.

In 1994, before the first long-term flight of female cosmonaut Ye. Kondakova, a 120-day hypokinesia test was conducted on female subjects, which made it possible to develop the recommendations necessary for her flight.

In 1999–2000, before the ISS went into full operation, the Institute of Biomedical Problems conducted a 240-day multifactor experiment, SFINCSS (Simulation of Flight of International Crew on the Space Station), involving participation by space agencies of Russia, the European Union, Canada and Japan, and also researchers from Austria, Germany, Norway, the United States, Sweden, and the Czech Republic. Twenty-seven volunteers from Austria, Germany, Canada, Russia, France, and Japan conducted more than 27,000 experimental trials and tested routine medical monitoring procedures. This experiment generated unique expertise in studying the interactions of an international crew and the long-term performance of a female crew member under conditions maximally approximating actual spaceflight conditions. Ground-based simulation experiments have generally preceded studies conducted on board unpiloted and piloted spaceflights, and their results have proven extremely useful in analyzing the data produced on spaceflights.

Research in Exobiology. Exobiology is the component of space biology that studies the presence, extent, and evolutionary characteristics of life in the Universe. Thus, exobiology asks such questions as, Where else in the Universe might life be found, and how would it be possible to establish its existence?

The awareness of the evolutionary interaction between the Universe and life led to extensive scientific search for such interactions in natural history. The issues addressed by exobiology may be divided into four areas; each corresponds to one of the major periods in the development of living systems:

- cosmic evolution, biogenic compounds;
- prebiological evolution;
- origin and early development of life;
- evolution of advanced life forms.

In its study of these areas, exobiology traces the paths leading from the beginning of the Universe to the major periods in the history of life. Additional aspects of exobiology address the study of compounds related to life and the search for life elsewhere in the Universe. Thus, exobiology develops through studying the occurrence and development of life on Earth, directly studying other planets and smaller celestial bodies of the solar system, and studying the rest of the Universe through observations from ground-based and orbital observatories.

The possibility of comparing biogenic products or extraterrestrial life forms with terrestrial forms is of enormous interest to biology. The problem of the existence of life beyond Earth is a component of one of the most important biological and philosophical problems—the origin and development of life in the Universe.

Studies in exobiology are conducted in two major areas: simulation of conditions in space or on certain planets and studies using robotic spacecraft. Vertebrates and higher plants are relatively sensitive to extreme factors, but microbial forms colonize virtually all possible ecological niches on Earth. There are organisms that can survive in a vacuum and those that can multiply under pressures up to 1300 bar and survive at pressures up to 20,000 bar. Some organisms can survive under exposure to ultraviolet radiation at a dose of 50,000 erg/mm² and ionizing radiation in doses of 2–4 Mrad. Many microbes can survive for long periods without access to external energy sources, without food, and with virtually no water.

As for the presence of extraterrestrial life forms (for example, on Venus or Mars), investigations conducted from robotic spacecraft have not yet yielded positive results. Nevertheless, all human activity in space occurs under conditions of biological quarantine. Safeguards must be taken to ensure that terrestrial life will not be exported to other celestial bodies and that possible extraterrestrial organisms are not brought back to Earth. The development of space biology and its accomplishments do not merely serve the goals of interplanetary travel and the human conquest of space. In the future, space biology will facilitate the construction of the most general biological concepts pertaining to the problem of life in its most general sense and the paths of evolution of life in the Universe.

Major Results. Biomedical experiments conducted on 85 crew missions ranging in duration from 23 to 438 days on space stations Salyut and Mir (USSR/Russia, 1971–2000), studies made on the Kosmos-110 specialized unpiloted biosatellites (USSR, 1966) and the 11 launches within the Bion research program (USSR, Russia, 1973–1997), and ground-based simulation experiments and investigations in general physiology and medicine have expanded our knowledge of the effects of spaceflight factors on living things, and in particular, on human beings. This has made it possible to improve continually the cosmonaut selection and training system and the means for medical monitoring of health to prevent the adverse effects of weightlessness and to provide postflight rehabilitation.

The major product of these efforts has been the creation of a unique system for preventing the adverse physiological effects of weightlessness, whose efficacy was convincingly demonstrated during space-station flights up to 1 year and longer (see Biomedical Support of Piloted Space Flight). On future spaceflights, requirements for maintaining cosmonaut health and supporting job efficiency and performance capacity will become more stringent as a result of increased space mission duration, requirements for enhanced EVA and assembly work, and the heightened complexity of research programs.

Understanding how living things react to extreme environmental factors and data on the limits of tolerance and endurance will make it possible to solve practical problems involved in designing and perfecting biotechnical systems and developing means and methods for increasing physiological tolerance to

Table 4. Results of Studies in Space Life Sciences

Space biology	Space physiology
Demonstration that weightlessness does not damage intracellular processes, cells, tissues, organs, or the organism as a whole	Definition of the major risk factors for piloted spaceflights
Demonstration that the development and growth of living things in weightlessness is generally normal without anomalous manifestations	Identification of the human functional systems most subject to alteration in space
Comprehensive study of the mechanism by which changes occur in various portions of the musculoskeletal system as a function of duration of exposure to weightlessness	Establishment of the main principles and stages of nonspecific and specific adaptive physiological changes in response to various extreme factors
Failure to identify any remote biological effects of space flight	Expansion of the understanding of the role of the gravitational factor
Identification of the time course of changes in vestibular functions in flight and quantitative description of the mechanisms of its adaptation to weightlessness	Study of the major mechanisms underlying regulation of functions and the formation and maintenance of homeostasis
Experimental demonstration that creation of artificial gravity on a spaceflight, through use of an onboard centrifuge, can prevent the development of a number of adverse physiological changes during spaceflight	Development of a rationale for approaches to predicting human tolerance and controlling defense reactions and reserve capacities under altered living conditions
Study of important mechanisms underlying the adaptation by various life forms to the effects of weightlessness and other spaceflight factors	Development of a system of medical measures and self-contained mobile devices for providing medical care under extreme conditions
Study of the combined biological effects of weightlessness, radiation, and other spaceflight factors	Design of means of rapid assessment of sanitary and hygienic conditions and ecological monitoring
Identification of the forms and limits of the dependence of living things on gravity	Refinement of requirements (including, toxicological ones) for an artificial living environment
Discovery of the principles underlying the evolution of microflora during long-term use of a space station, making it possible to evaluate medical, technological, and biospheric risks	Formation of a concept of the norm pertaining to various groups of healthy humans (profession, gender, age)
Study of the effects of heavy ions from galactic cosmic radiation on living things	Development of techniques for appropriate and corrective nutrition
Experimental demonstration during spaceflight of the utility of an electrostatic spacecraft shield against ionizing radiation	Expansion of the knowledge of human reserve capacities and human reactions to extreme situations

spaceflights, which is especially critical for supporting flights of human beings and their potential terrestrial fellow passengers (animals and plants).

One of the challenges faced by space biology is the study of the biological principles and methods for creating an artificial living environment in a spacecraft. Particular issues that need to be addressed, if this problem is to be solved, include finding living things that are promising components (subsystems) of a closed ecological system, identifying the combination of environmental factors and methods that will support the optimal population productivity and stability of such organisms, and simulating experimental biocenoses and investigating their functional characteristics and potential for use on spaceflights. The creation of an artificial closed ecological system—a living environment for humans on a spacecraft—would enable analysis and revision of the general biological significance and acceptability of traditional terrestrial living conditions and means of satisfying basic human needs. The main results of scientific studies conducted in space life sciences are provided in Table 4.

Among the important factors that have enabled human conquest of space, we can cite extensive coverage of the relevant problems, combining of traditional and innovative approaches to their solution, a close relationship between scientific exploration and the solution of applied problems, and collaboration between medical personnel and designers of space technology. All this has contributed to making space biology, physiology and space medicine sciences worthy of the twenty-first century.

BIBLIOGRAPHY

1. Gazenko, O.G., and A.I. Grigoriev. The major results of medical research in the USSR. *The World Space Congress*, Washington, DC, August 28–September 5, 1992.
2. Grigoriev, A.I., S.A. Bugrov, V.V. Bogomolov, A.D. Egorov, V.V. Polyakov, I.K. Tarasov, and E.B. Shulzhenko. Main medical results of extended flights on Space Station Mir in 1986–1990. *Acta Astronautica* 29 (8): 581–585 (1993).
3. Imshenetsky, A.A., D.V. Lysenko, and G.A. Kazakov. Upper boundary of the biosphere. *Appl. Environ. Microbiol.* 35: 135–155 (1978).
4. *Proc. Int. Symp. Non-Human Primate Research in Space*, Moscow, Russia, June 22–26. *J. Gravitational Physiol.* 7: 1 (2000).
5. Baranov, V.M. (ed.). *Simulation of Extended Isolation: Advances and Problems*. Slovo, Moscow, 2001.
6. Trofimov, V.I., A.N. Viktorov, and M.V. Ivanov. Selection of sterilization methods in planetary return missions. *Adv. Space Res.* 18 (1.2): 333–337 (1996).
7. Baranov, V.M., Ye.P. Demin, and V.A. Stepanov. Experimental studies of problems of habitability during long-term isolation in a pressurized enclosure. *Aviakosmicheskaya i ekologicheskaya meditsina* 31 (4): 4–7 (1997).
8. Gazenko, O.G., and Ye.A. Ilyin. *Evolution and Gravity (Issues of Evolutionary Physiology)*. Nauka, Leningrad, 1986.
9. Gazenko, O.G., A.I. Grigoriev, and A.D. Egorov. Medical research on long-term flights on Salyut 7-Soyuz-T. *Kosmicheskaya biologiya i aviakosmicheskaya meditsina* 24 (2): 9–15 (1990).
10. Gazenko, O.G., A.I. Grigoriev, and A.D. Egorov. From 108 minutes to 438 days... (on the 40th anniversary of Yu.A. Gagarin's flight). *Aviakosmicheskaya i ekologicheskaya meditsina* 35 (2): 5–13 (2001).

11. Grigoriev, A.I., S.A. Bugrov, V.V. Bogomolov, and co-authors. Review of the main medical results of the one year flight on space station Mir. *Kosmicheskaya biologiya i aviakosmicheskaya meditsina* 24: 3–10 (1990).
12. Zaloguyev, S.N., V.G. Prozorovsky, L.N. Kats, and co-authors. Structural and functional changes in bacterial cells under space flight conditions. *Doklady Akademii nauk SSSR* 278: 1236–1237 (1984).
13. Kovalenko, Ye.A., and N.N. Gurovsky. *Hypokinesia*. Meditsina, Moscow, 1980.
14. Yazdovsky, V.I. (ed.). Space biology and medicine. In: *Biomedical Problems of Space Flights*. Nauka, Moscow, 1966.
15. Mikhaylov, V.M. Hypokinesia as a risk factor under extreme conditions. *Aviakosmicheskaya i ekologicheskaya meditsina* 35 (2): 26–31 (2001).
16. Gazenko, O.G. (ed.). *Results of Biosatellite Investigations*. Nauka, Moscow, 1992.
17. Gazenko, O.G. (ed.). *Results of Medical Research on the Orbital Scientific Research Complex Salyut 6-Soyuz*. Nauka, Moscow, 1988.
18. Genin, A.M. (ed.). *Results of Scientific Research on Space Flights. Effects of Dynamic Space Flight Factors on Animal Physiology*. Nauka, Moscow, 1979.
19. Serova, L.B. *Ontogenesis of Mammals in Weightlessness*, series ed. O.G. Gazenko. Nauka, Moscow, 1988.

ANATOLY I. GRIGORIEV
DMITRY K. MALASHENKOV
Institute of Biomedical Problems
Russian Academy of Sciences
Moscow, Russia

SPACE PROGRAMS RELATED TO NATIONAL SECURITY

Introduction

The use of observation payloads on Earth-circling spacecraft, or satellites, was considered by U.S. military services as early as 1946. The Army Air Force asked Project RAND (Research on America's National Defense) of Douglas Aircraft Company, Inc, that later became The RAND Corporation, whether the use of space could be of military value. The investigators responded positively in several reports. In an early report in May 1946, RAND suggested that communications could be relayed from a low-altitude Earth-orbiting satellite (1). This was a decade before the launch of Sputnik. RAND issued a final report in 1954 (2), asserting that satellites that had the functions of communications, intelligence collection, navigation, and meteorology could be lofted on rockets being developed for intercontinental ballistic missiles (ICBMs). Curiously, the function of launch detection was omitted.

The use of spacecraft for reconnaissance would be a useful adjunct and, in many instances, a replacement for the use of aircraft that had been employed since World War I. More importantly, spacecraft payloads would provide a less

intrusive way to make observations over “denied” territory. To reinforce that concept, President Dwight D. Eisenhower declared an Open Skies Policy in 1955. The policy came about when President Eisenhower proposed to Soviet Chairman Nikita Khrushchev that mutual air surveillance be permitted to control nuclear weapons. The Soviets agreed to the principle, but they opposed the idea of inspection.

Today, military services use spacecraft for communications, navigation, and meteorology for military operations, in a manner similar to their civilian counterparts. In addition, military forces use spacecraft for detecting explosions and launching of rockets and aircraft and for earth observations through imagery and signals intelligence collection. Imagery can provide information on a target before a military strike and bomb damage assessment afterward. Electromagnetic intelligence (ELINT) provides the location and type of radar that indicates whether the radar should be targeted or should be avoided during a mission. The technical details of military spacecraft and their operations are kept secret. This is in contrast to civil and commercial spacecraft and their operations for which information is available to the public on request. The purpose of secrecy is obvious: military movements and information on which they are based should be hidden from enemy, or potential enemy, forces. In addition, the details of these spacecraft and the operations that collect “intelligence” information are also classified.

After World War II, both the Navy and the Air Force began to conduct research and development on space programs. The discussion that follows describes these space programs roughly in the order in which full scale development was authorized. For the sake of continuity, programs of the National Reconnaissance Office (NRO) will be discussed first.

Intelligence Information Collection

National Reconnaissance Office. During the years 1953–1961, President Dwight D. Eisenhower became increasingly concerned about the possibility of surprise nuclear weapon attack by the Soviet Union, a World War II ally turned adversary. Based on advice from trusted scientific advisers, President Eisenhower initiated a number of intelligence collection programs. For example, he authorized the development and production of the U-2 aircraft and its follow-on, the Oxcart (A-12), whose two-seat version became the better known SR-71. These aircraft were designed to collect imagery over the Soviet Union to evaluate its military capability and, indirectly, the intent of its leaders.

During these years, the U.S. Air Force began to develop spacecraft for imagery collection (SAMOS) and infrared rocket plume detection (MIDAS) (2), under a Weapon Systems program, WS-117L. In late August 1960, to limit access by the press, tighten security, and reduce possible objections from the Soviet Union about the United States development of a weapons system for space, President Eisenhower transferred the imagery program from the Air Force to a newly established covert organization that ultimately became the National Reconnaissance Office (NRO). His action changed the national priority. Intelligence data collection would first be devoted to strategic indications and warning of an attack

on the United States and second to tactical military operations. The new Secretary of Defense of the Kennedy administration, Robert S. McNamara, formally established the NRO, now composed of elements of the Air Force, Central Intelligence Agency, and the Navy, on 6 September 1961. As its first Director, he named Dr. Joseph V. Charyk, then the Under Secretary of the Air Force (2).

Ten years later, the NRO had evolved into a progressive organization responsible for research, development, acquisition, launch and operation of space systems for reconnaissance. The original concept was that a few very talented individuals would staff the organization, lay audacious plans, and guide it. They would depend on contractors to perform the research and development, produce the payloads, satellites, boosters, and in-orbit control vehicles. Launch and in-orbit control were to be conducted by teams of government and contractor personnel. Government teams were comprised of civilian and military personnel. This arrangement permitted rapid communication to take advantage of new technology, on the one hand, and on the other, to learn about operational difficulties that needed correction in later spacecraft and in in-orbit control techniques.

Organizationally, the NRO had a Director, a Deputy Director, and a small headquarters staff whose elements reported to a Director of the Staff. The field offices were organized as follows: Program A—U.S. Air Force, Secretary of the Air Force Special Projects (SAFSP), for some spacecraft programs and all launch vehicles; Program B—or CIA Reconnaissance Programs using both spacecraft and aircraft activities; Program C—U.S. Navy for spacecraft; and Program D—U.S. Air Force aircraft activities. These offices competed with each other in acquiring new technology and systems, sometimes constructively and sometimes not. The various projects within the NRO field offices were structured for maximum efficiency (stovepiping), not necessarily with commonality of architecture, subsystems, or components. The Gulf War in 1991, Operation Desert Storm, proved both the value of the NRO products and the difficulty of delivering them to the military field commanders. This is not surprising because the NRO was never intended to support military operations in the field.

The NRO delivered the collected photographs to the National Photographic Interpretation Center (now a part of the National Imagery and Mapping Agency) to extract intelligence from the imagery. The NRO delivered radar and communications intelligence data to the National Security Agency for signals intelligence, in general, but the CIA processed telemetry intelligence until about 15 years ago. The analyzed intelligence products were delivered to analytical intelligence organizations responsible for national estimates and to strategic and tactical military forces. Collection requirements for the NRO reconnaissance assets were developed by these users of the information.

Corona Operations: Imagery Collection. The first covert reconnaissance spacecraft was approved for full-scale development by President Eisenhower in 1958 and was called the Corona program (3–7). “Discoverer” was the overt name of Corona, but Discoverer had some missions of its own. After 13 previous tries, Corona Mission 9009 was successfully launched from Vandenberg Air Force Base on 18 August 1960, on a Thor Agena booster and second stage. This fourteenth attempted launch occurred on the very day that Gary Powers was being sentenced in Moscow for the U-2 shootdown. This Corona mission collected imagery

of one-half the area compared to that collected by all U-2 flights to date. This mission photographed 1.5 million square miles of the Soviet Union and East European countries. It photographed 64 Soviet airfields and 24 new surface-to-air missile (SAM) air defense sites. Ferret aircraft flying along the periphery of the Soviet Union and the Warsaw Pact countries had collected earlier detections and locations of SAM sites. This first successful mission of Corona impressed all those who viewed the photographs, from the analysts to President Eisenhower.

It should be noted in passing that the thirteenth attempted launch of Corona was also successful, but it carried diagnostic instrumentation rather than a camera and film load. The rocket boosters of that era did not have enough lift capability to carry both a photographic payload and sufficient instrumentation to measure the performance of the rocket during the launch phase and the spacecraft in-orbit.

The resolution of the panoramic KH-1 camera was 40 feet at nadir. The camera photographed from one side horizon to the other and slightly overlapped adjacent photographs in the direction of its travel trajectory. "Resolution" is the ability to separate and discern the narrowest two lines of equivalent width and spacing, assuming good contrast between, say, white lines and black background as in a parking lot. Resolution should not be confused with detecting the narrowest width white stripe on a black background in a parking lot. The resolution achievable from a spacecraft payload depends on the telescope of the camera, specifically, the lens aperture and the focal length. More importantly, it depends on the altitude of the spacecraft, on the (image) motion compensation (note, however, that it is the camera that is moving) and on the stability of the camera and the absence of vibration. For a film based camera, a simple calculation of the diffraction limit will yield a resolution that is better by a factor of about 10. The Corona camera went through a number of modifications which were numbered KH-2, -3, -4, -4A, and -4B, and the resolution ultimately improved to 5 feet. Stereoscopic capability was added so that analysts could view the targets in three dimensions. The U.S. Army Topographic Command, which later formed the core of the Defense Mapping Agency, could now produce terrain maps from Corona imagery.

Because a typical mission was only a few days long, there was a concerted attempt to increase the number of operating "days in orbit." The amount of film carried was increased and was eventually spooled into two capsules that were recovered separately. The amount of orbit adjust gas stored on board was increased to permit longer missions. Because a single satellite could not photograph any location on Earth immediately on demand, the NRO produced a term called "near real time." That meant that photographic coverage could be obtained only when the satellite made a pass over the target whose photograph is desired. Today, this term is also used to mean the time taken to process data that have been collected, but that was not the original meaning. To improve the daily coverage, more missions were orbited or "flown," and the result was that in the late 1960s, and early 1970s, the NRO dominated the number of U.S. space launches per year.

To recover on earth an exposed film capsule attached to the Corona satellite required a carefully orchestrated sequence of events. It began with separation of the capsule from the spacecraft, followed by firing retrorockets to decelerate the

capsule, appropriate stabilization for reentry, the deployment of parachutes to slow the fall through the sensible atmosphere, and snatching the assembly in mid-air by a crew flying a C-130 cargo aircraft based in Hawaii and equipped with a special harness.

A casual belief that once launched, a satellite will stay in that particular orbit, is actually incorrect. The satellite orbit is perturbed by the very unsymmetrical gravitational field of Earth. In geosynchronous Earth orbit, for example, there are only two stable points toward which all satellites tend to move. In addition, atmospheric drag changes the orbit of low-altitude satellites considerably in unpredictable ways. It is known that an atmospheric bulge is produced on Earth about 7 hours after a solar flare occurs. On one occasion, Corona passed over a ground control station one minute late, indicating an ephemeris uncertainty of 240 nautical miles!

In later years, Corona was flown at an altitude of about 100 miles at a somewhat retrograde inclination of about 110° . The last Corona mission occurred in 1972, and it was replaced by spacecraft that had greater capability. Because all space-qualified spacecraft had been launched, a working model was assembled, and given to the Smithsonian Institution (Fig. 1).

Operations for Corona were directed from an NRO Operations Center in a secure vault in the basement of the Pentagon. The Operations Center had detailed data on the dates that cloud-free photographs had been collected over targets and areas of interest. The users of the intelligence information specified the targets and frequency of collection as collection requirements. From these data, personnel at the Operations Center sent general collection directions by classified message to the Satellite Control Facility at what is now Onizuka AFB, named after the astronaut who was aboard the Shuttle Challenger when it exploded soon after launch. Onizuka AFB is located in Sunnyvale, California. From there, detailed instructions were sent to ground stations around the world. The ground station personnel loaded the instructions by microwave data links onto



Figure 1. Corona photographic satellite viewed from the side. Note on the left the two panoramic cameras for stereoscopic photographs and the film capsule on the right. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Corona for execution. The ground stations had cloud cover data from the weather satellite described later to influence the instructions to the film-limited Corona satellite.

Corona Products: Importance of Imagery—the Cold War, Arms Control, Weapons of Mass Destruction, and Verification. In the early days of the Cold War, it was important for the U.S. government to understand the military capability of the Soviet Union, as mentioned earlier. The ability to verify compliance with the Treaty became important as various treaties, such as the Strategic Arms Limitation Treaty I, were signed with the Soviet Union. That was to ensure the United States of compliance by the Soviet Union and also to confront them on practices that restricted verification, such as sheds to protect workers from weather conditions over missile silos and submarine pens. On a number of occasions, the Soviet Union confronted the United States about its practices to which they had objected, so the activities were symmetrical. The methods and technologies of collecting verification data came to be known as National Technical Means (NTM) in the arms control community.

Corona played a vital, but by no means the sole NRO, NTM role in monitoring the 1972 SALT I Interim Agreement on Strategic Missiles. By then, the NRO had other photoreconnaissance satellites in operation. Nevertheless, Corona's wide area search capability provided early detection of possible ICBM launch complexes under construction. Corona located the first Soviet ICBM complex at Yurya (Fig. 2), 500 miles east of Moscow (8). It was observed that one of the launch sites at this complex was dismantled as a result of the Interim Agreement.

In countries other than the Soviet Union, Corona (KH-3) Mission 9029 provided the first coverage of a Chinese nuclear test site near Lop Nor in December 1961 (8). In August 1964, the Director of Central Intelligence published a Special National Intelligence Estimate that assessed the likelihood that

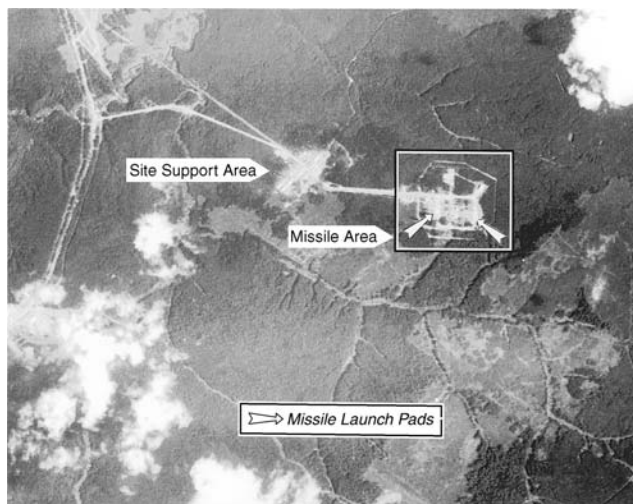


Figure 2. Corona photograph of ICBM complex at Yurya, USSR.

China would detonate its first nuclear device in 1964. Four days after the test of 16 October 1964, Corona took a photograph of Lop Nor (Fig. 3) that showed a crater from a nuclear blast.

After the Cold War ended, the need for imagery increased because of the requirement to monitor countries that had suspected weapons of mass destruction (WMD), understood to be nuclear, biological, or chemical (NBC) weapons. Today, such monitoring of possible production of NBC WMD in countries, such as Iraq, and of terrorist activities, such as in Libya and elsewhere is needed. Monitoring is needed also to detect and catalog means of delivering these weapons, especially ballistic missiles. Ballistic missiles are available, or under development, in 14 countries in the Middle East (southwest Asia), south Asia, east Asia, and South America (9). Delivery by even short-range ballistic missiles can threaten a substantial number of countries.

117 Operations: Meteorology and Weather Prediction. Weather prediction was important for the film-limited Corona satellite for two reasons: (1) to know before the imaging passes over a certain geographical area whether cloud cover would obscure the target and (2) also to know whether the film capsule could be recovered by the air crew over the Pacific Ocean in reasonably clear weather. Because of the civil applications of weather forecasting, it was agreed that the development of the meteorological satellite program would be the responsibility of the National Aeronautics and Space Administration (NASA) in conjunction with the Department of Commerce. This satellite program was to support both the civil and military communities. After delays in the development of a civil program, the NRO Director, Joseph Charyk, decided to proceed with an inexpensive, but high-risk weather satellite program (Program 417) to support Corona (10). The first satellite was launched on 23 August 1962 in time to support Corona Mission 1010. The 417 satellite successfully reported cloud cover over the Soviet Union on the following day. It operated at an altitude of 450 miles

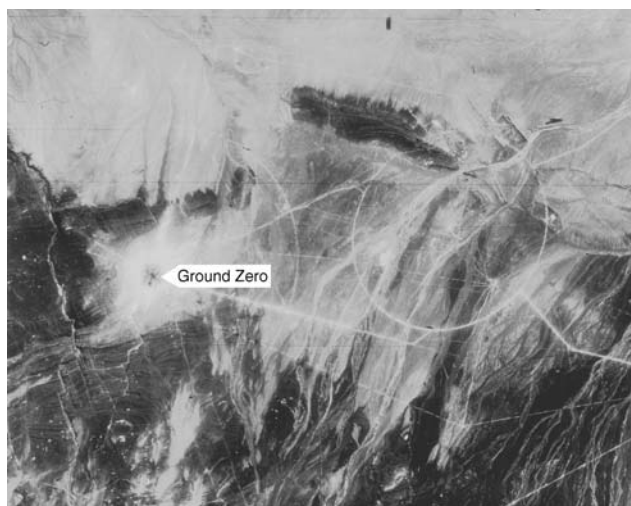


Figure 3. Corona photograph of nuclear test site at Lop Nor, China.

in a 96° retrograde orbit. By 1964, the NRO had four weather satellites in orbit while still awaiting the first National Operational Meteorological Satellite System. During the Cuban Missile Crisis, the NRO provided cloud cover information over both the Caribbean and the Soviet Union to U.S. officials. In 1969, it predicted clear weather over the Pacific Ocean in the Corona capsule recovery area, whereas the National Weather Service had predicted that clouds would obscure the recovery. The NRO decided to proceed with the mission and successfully recovered the capsule.

In 1971, Dr. John L. McLucas, the director of the NRO, decided to declassify the meteorological program because it had been producing weather photographs used in the Vietnam War that had been circulated widely. At that time, it was named the Defense Meteorological Satellite Program (DMSP). The Deputy Director of the NRO, Dr. F. Robert Naka (author of this article), resisted attempts by the contractor to add more capability to the satellite by arguing that a simpler system that met the needs of the NRO would be more reliable. The fundamental need of the NRO was to forecast cloud cover. That meant existing cloud cover and the direction of the winds and also measuring and calculating the vertical profile of the temperature and moisture to predict the formation of clouds. By the time the declassification process was completed and the organization was staffed for the use of the civil community in 1973, the DMSP (Fig. 4) and the existing low altitude civil program were essentially the same. The two programs were combined in the late 1990s.



Figure 4. DMSP weather satellite. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

GRAB Operations: Signals Collection. The second covert satellite reconnaissance program to be approved by President Eisenhower for full-scale development in August 1959 was the Galactic Radiation and Background (GRAB) program (11). He approved it for launch on 5 May 1960, just five days after the U-2 was shot down by the Soviet Union. It was successfully launched on a Thor Able rocket from Cape Canaveral on 22 June 1960 piggy-backed with the Navy's third Transit navigation satellite. GRAB (Fig. 5) was placed into a 373×748 km orbit at an inclination of 51.3° . The mission lasted 10 months and eventually ended when the spin stabilization was slowed down by Earth's gravitational field. Thus, it became the first operational satellite of the NRO, although not the first reconnaissance launch to be attempted. GRAB had two payloads: one to measure radar signals in the 3000 MHz S-band from the Soviet Union and the second to measure solar radiation (SolRad). The second mission was disclosed to the public. Both payloads were designed and produced by the Naval Research Laboratory. The ELINT payload successfully collected the signals from many more Soviet air defense radars than had been anticipated.

SUPPORT TO MILITARY FORCES

Launch Detection and Space Surveillance. The concept that a satellite-borne infrared sensor array could detect and report the launch of an ICBM and thereby provide 30 minutes warning was strongly supported by the military officers of the Strategic Air Command and the U.S. Air Force. The ground-based Ballistic Missile Early Warning System (BMEWS) could provide only 15 minutes warning, so the additional 15 minutes would permit launching more manned bombers from their air bases where they would be less vulnerable to a nuclear weapon burst. On the other hand, the civilian executives in the Pentagon were skeptical that the system could be made to work. More precisely, they were skeptical that the sensor array would have enough sensitivity and if it did, they thought that the background would produce false alarms.

Credit for pushing the idea of the Missile Detection and Alarm System (MIDAS) is given to Joseph J. Knopow employed by the Lockheed Aircraft Corporation in Van Nuys, California. MIDAS was the forerunner of the Defense Support Program (DSP, Figs. 6 and 7) (12). Both had passive infrared detectors

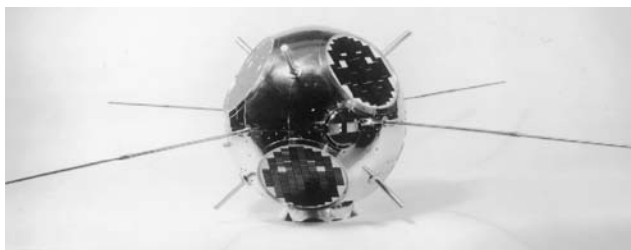


Figure 5. GRAB signals collection satellite.



Figure 6. DSP satellite. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

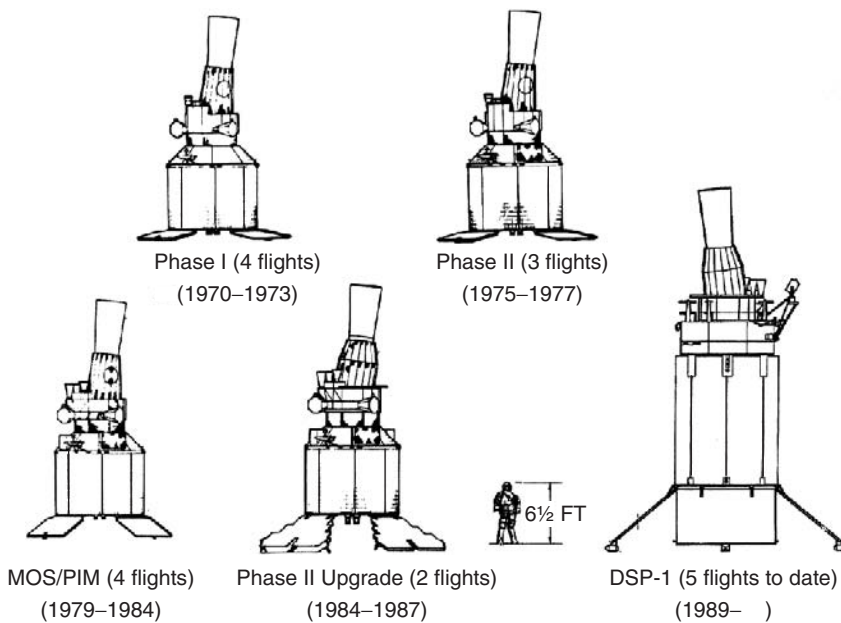


Figure 7. DSP configurations for various block changes.

that operated at a wavelength of $2.7\text{ }\mu\text{m}$ and could sense the radiation from a burning rocket plume. Here the similarity ends. MIDAS originally operated at a low Earth orbit (LEO) of 260 miles altitude whereas DSP, that is still operational today, operates at geosynchronous Earth orbit (GEO) at an altitude of 22,500 miles. There are even greater differences because the developers of MIDAS suffered from an immature technology that resulted in many launch failures coupled with in-orbit payload failures, even when the launch succeeded. The first attempted launch on 26 February 1960 ended in failure.

In June 1956, Lockheed Missiles and Space Division was selected as the prime contractor. MIDAS never left the research and development phase, so that there was no full-scale development decision. Rather, the emphasis was on gathering sensitivity data and on making the system work. MIDAS had four series; each had two to four spacecraft. Series IV was intended as a prototype but was cancelled. The spacecraft were numbered sequentially irrespective of the Series. Not until MIDAS 7, the second of Series III, launched on 9 May 1963, was total success achieved. The spacecraft operated for 6 weeks. The capability of the concept was 50 kW/srad , and the background produced no false alarms. Series I and III had Aerojet sensors, and those of Series II were built by Baird Atomic. Series III had more capability than Series II, and Series II more than Series I. Series I had 27 detectors spinning at 2 rpm, whereas Series III had 184 lead sulfide detectors spinning at 6 rpm at an altitude of 2000 nm.

In the late spring of 1963, Project Forecast was convened to recommend promising new concepts and technologies for the Air Force. General Bernard Schriever and Dr. Allen Puckett led the study. The success of MIDAS 7 encouraged expansive thinking. The Surveillance Panel recommended that it was time to think of putting MIDAS technology into geosynchronous Earth orbit (GEO) (13). In January 1968, a study, "Surveillance of Objects in Space in the 1970s" was initiated jointly for General James Ferguson, Commander Air Force Systems Command and for General Arthur Agan, Commander Aerospace Defense Command, and directed by the author (14). The Study recommended detecting ICBM launches by using a DSP-like system that was called High Altitude Surveillance Platform (HASP) and passive Long Wavelength Infra Red detector system ($8\text{--}14\text{ }\mu\text{m}$) in LEO called Low Altitude Surveillance Platform (LASP) to catalog objects in space and differentiate them from nuclear warhead reentry vehicles and decoys. During the Strategic Defense Initiative (SDI), the LASP concept was changed and called the Space Surveillance and Tracking System, now called Space Based Infra Red System (SBIRS) Low. SBIRS Low follows the SDI concept of cueing from DSP (now called SBIRS High) and tracking those objects, so it has a narrow field of view that does not permit the broad space surveillance proposed for LASP.

Communications. This section contains a brief discussion of military space communications. There is a more extensive discussion of communication satellites elsewhere in this Encyclopedia. This section first describes the differences between military and civil communications systems, then three military satellite communications programs, roughly in the order in which they were approved for development and acquisition, although the reader will note that the programs ran concurrently. They are the Defense Satellite Communications System for long-haul communications, the Tactical Satellite Communications System for tactical communications, and Milstar for strategic communications.

Military communications that employ Earth satellite relays are similar to their civil counterparts, but military satellites have one distinctive difference. Military communications must be able to operate before, during, and after a military conflict. They must be designed to withstand intentional enemy interference, or "jamming," as well as unintentional interference. The uplink from ground to satellite is susceptible to enemy jamming because a determined adversary can generate considerable power and employ a large antenna at any location within the uplink coverage area of the beam on Earth. The countermeasure for this is to have a large bandwidth system that employs frequency-hopping, or spread spectrum signals, to force the adversary to broadband noise jamming that dilutes the effort. This countermeasure favors higher operating frequencies because radio-frequency component bandwidths are generally 10% of the operating frequency. The downlink is somewhat less susceptible to jamming because it is the ground or airborne receiving terminal that must be jammed, so that main beam jamming is usually not possible. Side-lobe jamming requires a factor of about 1000 (30 decibels) in power density greater than that needed for main beam jamming. On the other hand, satellite power is limited by what can be generated in space. To offset this weakness, multiple narrow beams are employed in the satellite that again favors higher operating frequencies. Hence, there is a trend from ultrahigh frequency (UHF) in the military band from 225–400 MHz toward extremely high frequency (EHF) in the military band from 20,000–44,000 MHz.

There is also a need to transmit classified messages either for military operations or other sensitive data such as intelligence that will require message encryption. In the civil community, there is a similar need to transmit banking, credit, and currency data, but much of that traffic is carried by fiber-optic cable. On the other hand, military communications, must be available worldwide where often fiber-optic cable communications are not available, such as in much of Africa. Nevertheless, in 1994, the Air Force Scientific Advisory Board recommended that the Air Force make more use of fiber-optic cable communications instead of satellite communications, whenever it is available (15). In many instances, the life-cycle cost of fiber-optic communications is less, even for trans-oceanic communications.

The fact that commercial companies had laid the fiber was not a problem for the military services. The Department of Defense (DoD) has often leased private networks for its classified operational message traffic and used commercial networks for its routine traffic.

A similar situation occurred in using satellite communications. As military satellite communications developed, there was a tendency to put all military traffic on dedicated DoD networks. However, that capacity was found insufficient during the Vietnam War, so routine traffic was again carried by commercial satellite carriers. That decision reestablished the policy that had been in effect when there were only landlines for communications.

Using those facts as preamble, the first military satellite communications project was called Advent and was assigned to the U.S. Army for development. There were prior activities that are described in the literature (16). Advent suffered through the same research and development problems that plagued space programs in general: immature technology led to delays and cost overruns. That in turn led to criticism by senior executives, such as the Secretary of Defense,

and by the members of Congress and caused program cancellations. In the case of Advent, Defense Secretary McNamara cancelled the Program on 23 May 1962. Before that date in May 1960, the strategic, as opposed to tactical, communications of the three services were combined into the Defense Communications System (DCS) and were assigned to the newly created Defense Communications Agency (DCA). On 15 July 1964, Secretary McNamara approved a system for full-scale development that The Aerospace Corporation had been studying for the Air Force (17,18) and thereby approved a dedicated military system that the Air Force had favored over a commercial joint use civil and military system. A joint use system is not necessarily ideal because the civil users are always suspicious of any military incursions into their domains. Conversely, the military are uneasy about the reliability and security of civil systems.

The new system was called the Initial Defense Communications Satellite Program (IDCSP). The satellites were to weigh 100 pounds each and would be launched up to eight at a time in near synchronous circular orbits by a Titan IIIC booster. The Air Force was given responsibility for the satellites and boosters, and the Army was given responsibility for the ground terminals. The first seven satellites were launched on 16 June 1966, operating in the super high frequency (SHF) band of 8–9 GHz. A total of 26 satellites was successfully launched by 13 June 1998. Each satellite had a 24-face polyhedron that contained a total of 8,000 solar cells. This was sufficient to power either 11 tactical quality or five commercial quality voice circuits or 1550 teletype symbols per second. Alternatively, each satellite could relay about 1,000,000 bits of digital data per second. DCA declared the system operational before the launch of the last group of satellites and changed the name to Initial Defense Satellite Communications System (IDSCS). The satellite's mean time before failure turned out to be 6 years although the design life had been 3 years. As the shortcomings of IDSCS were remedied by advanced systems, it became known as DSCS I (Fig. 8).

A DSCS II satellite constellation was to consist of four satellites in geosynchronous orbit plus two orbiting spares. Each satellite was larger than DSCS I at 2.7 m in diameter and 4 m high, when antennae were extended, and weighed 590 kg. It was dual spun and could support 1300 two-way voice channels or 100 million bits of digital data per second. The first pair was launched on 2 November 1971. The remaining history is somewhat spotty because of the now usual booster failures and also payload failures. By the early 1980s, however, the constellation had fulfilled the DoD's need by linking 46 strategic ground terminals and 31 mobile forces ground terminals and also served 52 terminals of the Diplomatic Telecommunications System. In addition, DSCS B4 launched 13 December 1973 lasted four times its design life and was turned off on 13 December 1993.

DSCS III finally came close to the operational needs of the armed forces by having separate electronically switched SHF multiple beam antennae for transmission and reception, as well as UHF and SHF Earth coverage antennae. Antijam capability was provided by the antenna diversity. Naturally, each satellite was heavier and larger, but its main difference was that it was three-axis stabilized.

The tactical system or TACSATCOM at ultrahigh frequencies (UHF) in the 225–400 MHz band is the responsibility of the U.S. Navy. It is worth noting that



Figure 8. DSCS satellite. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

all military aircraft are equipped with UHF radios, so this a very important program, even though aircraft may not be equipped to communicate with satellites. Of the three systems described here, UHF radios are the most susceptible to jamming.

The triservice Tactical Satellite Communications Program 591 was established in 1965 by the DoD to enable the Army, Navy, and Air Force to determine the potential usefulness of UHF satellite communications. The Lincoln Laboratory of the Massachusetts Institute of Technology was chosen to develop the experimental satellites for the test program (19,20). Lincoln Experimental Satellite (LES) 5 was to be launched as soon as possible, and LES-6 was to be launched a year later incorporating improvements. LES-5 was launched on 1 July 1967 on a Titan III-C and LES-6 on 28 September 1968. LES-5 was placed into a subsynchronous equatorial orbit and drifted eastward at 33° per day. The tests of Army, Navy, and Air Force ground, sea, and air units met their performance goals, and the system was put into production. To complete a worldwide communications system, a number of satellites have been put into orbit over the years. They are TACSAT in 1969, three Gapfiller/GAPSATs in 1976, six FLTSATCOMs in 1978 to 1989, four LEASATs in 1984, and 10 UHF Follow-on (UFO) satellites in 1993 to 1998. UFO F-11 is scheduled for launch in year 2004, the Mobile User Objective System is in program definition, and is expected to be operational in 2007.

The UFO program has more than an improved UHF capability. Flight vehicles F-1 through F-3 had an SHF payload as well, an UHF F-4 through F-7 had EHF in addition, and F-8 through F-10 have a Global Broadcast System (GBS) instead of the SHF payload. The GBS is to broadcast intelligence information to deployed U.S. Forces (15). Thus the UFO satellites have an expanded capability to support, simultaneously, multiple users in several military bands.

The strategic system, the responsibility of the Air Force, was called Stratcom, now Milstar. It operates at extremely high frequencies (EHF) of

20 GHz (for the downlink) and 44 GHz (for the uplink) and is closer to an ultimate system in that it is robust. Its wide bandwidth would make jamming very difficult and smaller ground terminals could be supported. However, the issue of backward compatibility has caused Milstar to carry also both SHF for long haul communications and UHF payloads for tactical users. To support this wide array of users, a 60 GHz cross-link is available between the geosynchronous altitude satellites. The result was a Milstar I satellite (Fig. 9) weighing 10,000 pounds that required a Titan IV booster to launch it rather than a medium launch vehicle. To keep the weight down, the highly protected individual channel satellite capability has a low data rate of only 2.4 kbps. Milstar II will increase that capability to 1.544 Mbps. There are currently two Milstar I satellites in orbit, F-1 launched on 7 February 1994 and F-2 on 6 November 1995. The first launch of Milstar II, F-3, failed in May 1999. The second launch, F-4, is planned during 2001.

Navigation. This section describes the development of satellite-borne navigation systems by the Department of Defense. There is a more thorough discussion elsewhere in this Encyclopedia. This section describes the differences between military and civil navigation systems and the development of the Global Positioning System (GPS) by the Department of Defense. It also describes the measures taken to ensure that use of GPS by the enemy in the war zone will be ineffective and enemy attempts to disrupt U.S. use of GPS can be frustrated.

There are two major differences between military and civil systems that should be discussed. The first is jamming, attempted deliberate interruption of the system by an enemy in a war zone. The second is the need for our forces to interrupt civil GPS receivers being used by the enemy in the war zone, a process called "denial." The denial process should not disrupt civil uses of GPS in a country adjacent to the war zone.

Accurate position location and navigation are at the heart of many activities, not only for both the military and civil communities. The navies and

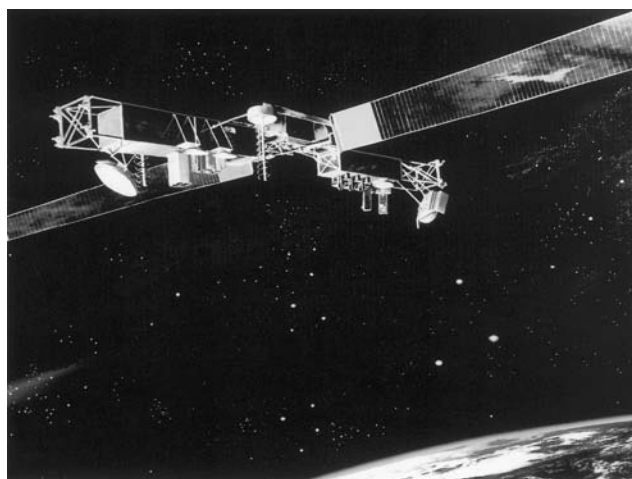


Figure 9. Milstar satellite. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

merchant marines of the nations require accurate navigation to reach their desired locations. It is not surprising, therefore, that the first navigation satellite system, Transit, was put into operation by the U.S. Navy in 1964 and is still in use today (21). This satellite was designed by personnel at the Applied Physics Laboratory of the Johns Hopkins University. A follow-up program called Timation, built by the Naval Research Laboratory, proved the ability to operate atomic clocks in orbit. That was used as a system concept for the Global Positioning System. Program 621B was a study conducted by the Aerospace Corporation in 1964 for the Air Force Space and Missile Systems Organization. In 1972, the best characteristics were assembled for a new system called NAVSTAR Global Positioning System. The system was developed and built for the U.S. Air Force that operates and controls it from a ground station at Schriever Air Force Base, near Colorado Springs, Colorado. The space portion is comprised of 24 satellites, 4 each in 6 planes. Each satellite knows its own ephemeris and the time, so that it knows its position at all times.

Military and civil position locations require a number of similar and essentially identical activities but also serve vastly different ones. Both communities would like to land aircraft on a runway under conditions of very poor visibility, preferably automatically without human intervention. At the other extreme is the military desire to lay bombs on military targets very accurately, employing small bombs that destroy the targets but leave adjacent structures and people unaffected. The recent bombing of buildings in Belgrade, Serbia, by forces of the North Atlantic Treaty Organization is a case in point. This great accuracy was made possible by GPS. This pinpoint accuracy contrasted sharply with what was possible during the Vietnam War, when it was impossible to destroy a bridge using conventional bombs dropped from B-52 bombers from high altitude.

As implied by the last Navy system, Timation, the exact time that is carried by all elements of the navigation system, is the most important factor. All GPS satellites (Fig. 10) carry either rubidium or cesium clocks, or both. These have an accuracy of about 10 nanoseconds that corresponds to an accuracy of 10 feet. The U.S. Naval Observatory has installed a hydrogen maser clock that has an accuracy of one picosecond at the ground control station at Schriever Air Force

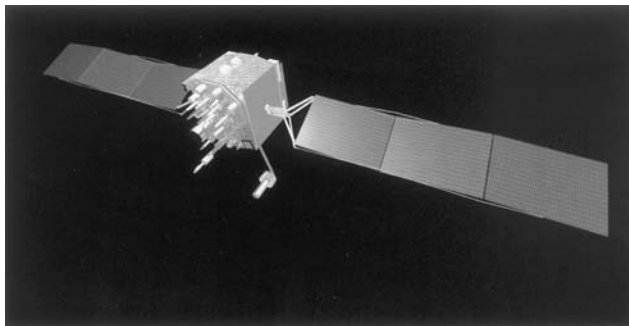


Figure 10. GPS satellite. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Base. If this accurate time could be transmitted routinely to the satellites for retransmission and a system error of a factor of 100 were adopted, the accuracy of the entire system could be improved to one inch!

BIBLIOGRAPHY

References for the recently declassified portions of this article are primarily government documents written by government employees or authorized by the government. These references are more likely to be accurate than many of the secondary articles and books written on these subjects.

1. Preliminary design of an experimental world-circling spaceship. Report No. SM 11827, Douglas Aircraft Co., USA, 2 May 1946.
2. Lipp, J.E., and R.M. Salter (eds). Project feedback summary report. R-262, Volume I. The RAND Corporation, USA, 1 March 1954.
3. Hall, R.C. Missile Defense Alarm, The Genesis of Space-Based Infrared Early Warning. National Reconnaissance Office History Office, USA, July 1988.
4. Oder, F.C.E., J.C. Fitzpatrick, and P.E. Worthman. The Corona Story. National Reconnaissance Office, USA, December 1988.
5. Ruffner, K.C., et al. CORONA: America's First Satellite Program. CIA History Staff, USA, 1993.
6. Wheelon, A.D. Corona: The First Reconnaissance Satellites. *Phys. Today* 50 (2): 24–30 (February 1997).
7. Hall, R.C. Postwar strategic reconnaissance and the genesis of CORONA. In D.A. Day, J.M. Logsdon, and B. Latell (eds). *Eye in the Sky*, The Smithsonian Institution Press, Washington, DC, 1998.
8. McDonald, R.A. (ed.). *Corona Between the Sun and the Earth*. American Society for Photogrammetry and Remote Sensing, USA, 1997.
9. Nolan, J.E., and A.D. Wheelon. Third World ballistic missiles. *Sci. Am.* 34–40 (August 1990).
10. Haines, G.K. The National Reconnaissance Office: Its Origins, Creation, and Early Years. National Reconnaissance Office, USA, 1997.
11. Day, D.A. Listening from above: The first signals intelligence satellite. *Spaceflight* 41 (8): 338–346 (August 1999).
12. Hall, R.C. Missile Defense Alarm. National Reconnaissance Office, USA, July 1988.
13. Report of the Surveillance Panel, Project Forecast. U.S. Air Force, USA, 1963.
14. Naka, F.R., et al. Surveillance of Objects in Space in the 1970s. Air Force Systems Command, USA, 1968.
15. Naka, F.R., et al. Communications Technology Options for Global Air Operations. Air Force Scientific Advisory Board, SAB-TR-94-03, USA, February 1995.
16. Spiers, D.N., and R.W. Surdevant. From Advent to Milstar: The U.S. Air Force and the challenges of military satellite communications. In A.J. Burriga (ed.), *Beyond the Ionosphere: Fifty Years of Satellite Communications*. The NASA History Series, USA, 1997.
17. Military communications satellites. In The Aerospace Corporation, *Its Work: 1960–1980*. The Aerospace Corporation, USA, 1980, p. 50.
18. Black, R.L., and W.L. Pritchard. Synchronous Communication Satellite Study, Summary Report, Vol. I, Purpose, System Data & System Studies. The Aerospace Corporation, USA, 1963.

19. Ward, W.W., and F.W. Floyd, Thirty years of space communications research and development at Lincoln Laboratory. In *Beyond the Ionosphere*. Air Force Systems Command, USA, 1968.
20. Martin, D.H. Communication satellites 1958 to 1986. The Aerospace Corporation, USA, October 1984.
21. *The Global Positioning System, A Shared National Asset*. National Academy Press, USA, 1995.

F. ROBERT NAKA
CERA, Incorporated
Concord, Massachusetts

SPACE RADIATION

Introduction

According to NASA's Critical Path Roadmap (1), the most critical biomedical risks of space habitation are (1) carcinogenesis caused by radiation, (2) loss of bone mass or density, (3) poor psychosocial adaptation, and (4) clinical manifestations of trauma or acute medical problems. Because of the unique problems and the potential long-term consequences, radiation is frequently considered the most serious of these for long-duration space missions, and cancer is the major risk followed by damage to the central nervous system. Radiative hazards in space have been discussed in depth by a Task Group of the National Research Council along with overall research strategies (2).

The term radiation in the present context refers specifically to directly and indirectly ionizing radiation, including X rays, gamma rays, neutrons, ions, and other somewhat esoteric particles such as muons and pions. Radiation is a natural part of our environment on the surface of Earth. We are constantly being bombarded with radiation from our own Sun and from outer space. In fact, X rays and gamma rays are part of the electromagnetic spectrum to which we are exposed daily that includes infrared radiation (heat), visible light, and ultraviolet radiation as well as higher energy X rays and gamma rays. Radioactive elements also are naturally present in essentially all of the materials that surround us as well as those of which we are composed. Biological organisms in general, and humans in particular, have adapted to the normal radiative environment at Earth's surface and certainly have even benefited from the radiation. At the same time, those radiations also initiate biological damage ranging from sunburn to cancer. Earth's atmosphere and its magnetic field act as excellent radiative shields by reducing the levels of radiation. As we go farther away from Earth's surface into orbit or toward outer space, the radiative level generally increases in intensity and changes in the types and energies of the radiation. It is known that the radiative dose rates for personnel in space are significantly higher than on Earth's surface. We have had little experience with regard to the biological consequences of healthy people exposed to these types of radiation for protracted periods of time. Our best estimates indicate, however, that the additional

exposure to higher energy radiation beyond that on Earth’s surface or above the present allowable limits for radiation workers could result in significant risks of cancer or other serious diseases.

Because of higher level of exposures and the potential for long-term negative consequences, radiation has been categorized as the worst biomedical risk of long-duration space missions. At the same time, using an appropriate programmatic strategy and focused research, the risks of radiation in space should be reduced to levels comparable to or less than those of other hazards.

Radiative Fields

Individuals in orbit around Earth or on space missions are subjected to three major sources of radiation: (1) particles trapped in Earth’s magnetic field (4), (2) radiation emanating from solar events, and (3) galactic cosmic rays. Each of these radiative fields has unique distributions of particle types and energies at fluence rates (number of particles per unit area) of varying intensity, but generally higher than those at Earth’s surface.

Trapped Radiation

Electrons. When subjected to an external magnetic field such as that surrounding Earth, charged particles spiral around the magnetic field lines. Thus charged particles, particularly protons and electrons, accumulate in significant numbers around Earth along the magnetic field lines. Figure 1 shows a typical frequency distribution of the number of electrons as a function of their energy (5). Because of the relatively low energies of these electrons, they are absorbed by spacecraft or space-suit materials and are not likely to represent a major source of exposure for personnel in Earth orbit.

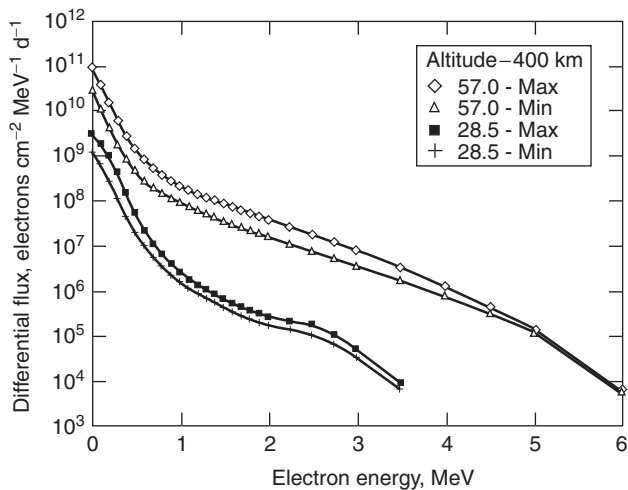


Figure 1. Trapped-belt electron spectra at 400 km for solar minimum and maximum inclinations of 28.5° and 57° using AE-8 environment model (see Reference 5). [Reprinted (or adapted) from NCRP Report (or Commentary) No. 132, copyright year (1999), with permission from the National Council on Radiation Protection and Measurements.]

Trapped Protons. A typical distribution of protons trapped in Earth's magnetic field (5) is shown in Fig. 2. Because of the higher energy and higher mass of these particles, they can deliver a significant dose to personnel who spend long times in orbit.

Solar Events. The Sun, operating largely as a result of nuclear fusion, is the major source of energy in our solar system. Although the rate of energy from the Sun is relatively constant, anomalous activities arise locally and stochastically, and these events can be major sources of radiation, especially higher energy protons. Figure 3 shows distributions of protons emanating from an exceptionally intense solar event as a function of time and the energy of the protons (6). Note that the fluence was delivered during an extended period of several hours. These protons can penetrate significantly and can produce substantial biological damage if left unchecked. They have, however, relatively low energies, so appropriate shielding can significantly reduce the doses and, presumably, the biological consequences. Additionally, although the occurrence of a specific solar event is somewhat stochastic, the number of such occurrences per unit time is cyclic, and the period is about 11 years. So the cyclic nature of solar events is something that may be used to minimize exposure. However, the fluence rate of galactic cosmic rays actually decreases during maximum solar activity, so the best scenario for minimum exposure is not necessarily during solar minimum.

Galactic Cosmic Rays

Galactic cosmic rays generally originate from activities within our galaxy during millions of years. As a group, these particles have the most energetic radiation and, therefore, are the most penetrating and the most difficult to shield. They

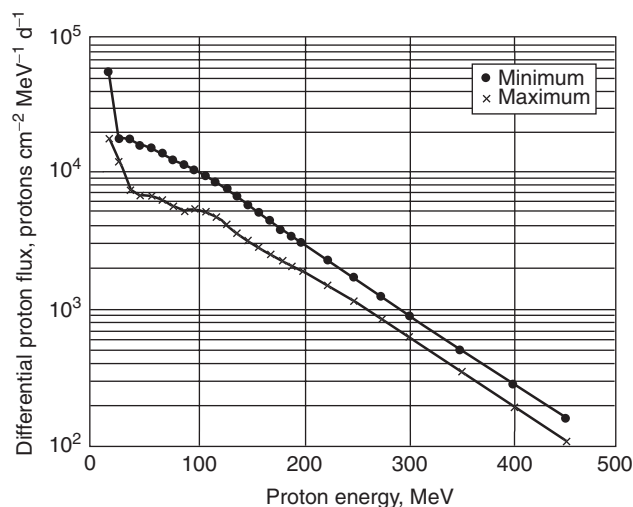


Figure 2. Trapped-belt proton spectrum for the ISS orbit (51.6° inclination, 470 km altitude) using AP-8 models for solar maximum and minimum (see Reference 5). [Reprinted (or adapted) from NCRP Report (or Commentary) No. 132, copyright year (1999), with permission from the National Council on Radiation Protection and Measurements.]

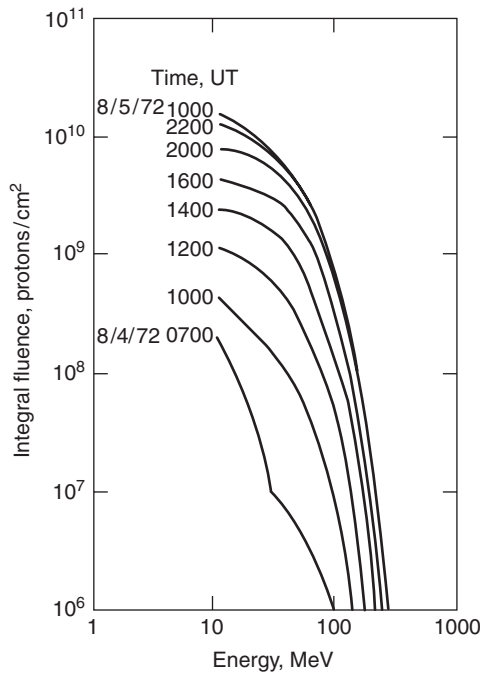


Figure 3. The proton fluence as a function of time and energy of a solar event in August 1972 (courtesy of National Council on Radiation Protection and Measurements).

consist primarily of protons and, to a lesser amount, helium ions; only a small fraction of ions has higher atomic numbers. The probability per logarithmic interval in kinetic energy per nucleon, $kEf(E)$, as a function of the kinetic energy E for four ions is given in Fig. 4. These particles have the highest energies of space radiation and have modal values of about 1 GeV (1 GeV is 10^9 electron volts, where an electron volt is the energy gained by one electron accelerating through a potential of one volt). The energetic ions with atomic numbers greater than that of helium are frequently called HZEs, i.e., heavy ions (H) with high atomic numbers (Z) and high energy (E).

Radiation on the Moon and on Mars

The radiative levels on the surfaces of both the Moon and Mars are reduced by their partial directional shielding and, in the case of Mars, the shielding provided by the Martian atmosphere. In both cases, the radiative fields are expected to be between those observed on Earth’s surface, where its thickest atmosphere and greatest mass are, and those of deep space. It is estimated that radiative doses on the Moon are about half as intense as those in outer space (7). The same estimates for the Martian surface are significantly lower than in free space or on the surface of the Moon. Although higher than at Earth’s surface, they could be within acceptable ranges for radiation workers, especially if the existing regolith were used for shielding.

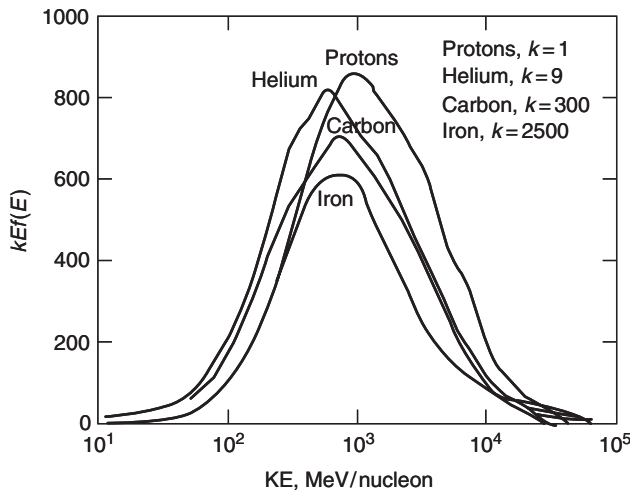


Figure 4. The differential probability per logarithmic interval in kinetic energy per nucleon, $kE/f(E)$, as a function of the kinetic energy E for four ions. $f(E)$ is the linear differential probability for a particle of kinetic energy E . k is the scaling factor given in the legend. In this representation, the area under the curve between two values of E is proportional to the number of particles in that interval (courtesy of National Council on Radiation Protection and Measurements).

Estimation of Doses from Radiation

During the last two decades, our knowledge of the types of individual particles present in the different radiative fields to which astronauts are likely to be exposed and their variation in time has improved considerably as a result of dedicated research efforts throughout the world. Concurrently, data to determine the doses to which the astronauts would be exposed were also obtained. The net result is that estimates of exposure levels are significantly more accurate, compared with earlier values, as are the corresponding doses for likely situations in space.

Terminology. The dose D is the energy deposited by radiation per unit mass of the absorbing medium [1 joule/kg = 1 gray (Gy) or 100 ergs/g = 1 rad]. It is sometimes called physical dose or absorbed dose to differentiate this quantity from other related quantities such as dose equivalent.

The relative biological effectiveness (RBE) is the ratio of the dose from a reference radiation, usually X or gamma rays (γ), to the dose from the radiation being studied to produce the same level of biological response ($\text{RBE} = D_{\text{reference}}/D_{\text{studied}}$).

In health physics, the term radiative quality Q is used for risk assessments and to establish the relative risk from different types of radiation or different energies. The unit for radiative quality is sieverts per gray (Sv/Gy). Analogous to the RBE, Q is a factor by which the physical dose is multiplied to obtain the dose equivalent to that of the reference radiation, usually X or gamma rays. The RBE is generally used as a research quantity to express relative effectiveness, whereas radiative quality is more often used in a regulatory manner for risk assessments.

The dose equivalent D_{EQ} , then, is product of the physical dose and the quality factor for a specific type of radiation ($D_{\text{EQ}} = Q_{\text{F}}D$). Thus, it is the dose of

the reference radiation which produces the same level of risk as the actual physical dose. The units are sieverts (Sv). The definition of dose equivalent intrinsically assumes that the physical properties of the dose D and the biological properties of the radiative quality Q are independent and multiplicative. Therefore, the relationship, further assumes that dose equivalent is linear with dose. None of these assumptions is necessarily true, and, therefore, Q is a function of dose, irradiative time, and other parameters.

Differences in biological response or risk of diseases for the same absorbed dose arise in part because particles of different energies, charge, or mass deposit energy differently as they traverse a given path length through the material. To quantify those differences, we define linear energy transfer (LET or L) as the energy lost by the particle per unit path length (MeV/m or keV/ μ m).

Dose and linear energy transfer are macroscopic quantities in the thermodynamic sense and, as such, apply only to large masses in equilibrium conditions. Many relevant biological processes take place at the cellular or molecular level. For these conditions, quantities such as dose or LET should not be applied because major errors result in some cases when this is inappropriately done. An entire field known as microdosimetry describes these microscopic and submicroscopic quantities and processes (8–11). To describe the stochastic process analogous to the macroscopic quantity LET defined in the previous paragraph and to differentiate between the macroscopic and microscopic processes, we define the quantity lineal energy y as the microscopic energy deposited per unit path length in units of keV/ μ m.

Doses from Environmental Sources. Radiation is an intrinsic and natural part of our environment. Radiation from the Sun is the major source of energy for Earth and has many positive consequences. At the same time, this same energy can produce undesirable consequences ranging from sunburn to cancer. Nevertheless, because we generally assume that average levels of exposure from natural sources are acceptable, these levels represent a good baseline for comparison. Table 1 (12) lists typical sources of radiation and their contributions to exposures of the U.S. population, at a total dose equivalent of about 0.4 cSv. The actual dose is a strong function of elevation and location and can easily vary

Table 1. Recommended Limits on Dose Equivalent

	ICRP	NCRP
Occupational		
Annual	5 cSv ^a	5 cSv
Lifetime (cSv)	—	Age ^b
Emergency	10 cSv	10 cSv
Public		
Annual (continuous)	0.1 cSv	0.1 cSv
Annual (occasional)	0.5 cSv	0.5 cSv

^a1 cSv = 100 crem = 1 rem.

^bAge in years.

[Reprinted (or adapted) from NCRP Report (or Commetary) No. 132, copyright year (1999), with permission from the National Council on Radiation Protection and Measurements.]

from this value by factors of two or more. Naturally occurring radon gases account for more than half of the dose equivalents, and cosmic radiation and other natural sources account for about a quarter of the dose equivalent. Medical applications and other man-made sources account for less than 20% of typical exposures.

Because the annual dose from natural sources hovers at about 0.3 cSv, it is not coincidental that the recommended upper limits for dose equivalents to the general public are typically between 0.5 and 0.1 cSv. Recommendations of two major institutions, the National Council on Radiation Protection and Measurements (NCRP) mentioned earlier and the International Commission on Radiation Protection (13) are compared in Table 2 (adapted from Reference 12). Occupational limits are generally allowed to be about a factor of 10 higher than those for the general population based on the strong stipulation that exposures should be kept *as low as* reasonably achievable, known as ALARA (14).

Most of our data for medical complications that result from exposures to radiation higher than the natural background come from either medical diagnostics or treatments, accidental exposures, or observations of survivors of nuclear explosions (15). Several decades of work have gone into medical evaluations and analyses of these latter individuals, and these data form the major source of information about risks from environmental exposures. These data must be used with care, particularly in space applications, because the atomic bomb exposures were acute exposures of a specific population under extreme conditions. Other human results along with animal and cell studies have been used to extend these results to estimate risks for humans in space (see Table 3). Two excellent reports by the National Council on Radiation Protection and Measurements go into details of both the analyses and the application to radiative protection in space (7,12). For the present discussion, the most important single quantity for comparative purposes is the overall risk per unit dose equivalent. The value, even of that quantity, varies with gender, age, and biological end point. Generally, cancer is considered the single most likely risk from radiation exposures in space. The overall probability that an individual will develop cancer in the United States is

Table 2. Annual Effective Dose Equivalent^{a,b}

Source	Persons exposed	Avg. annual dose equiv., cSv	Avg. annual dose equiv. In US pop., cSv ^a
Natural sources:			
Radon	230,000,000	0.2	0.2
Other	230,000,000	0.1	0.1
Occupational	930,000	0.23	0.0009
Nuclear fuel cycle	—	—	0.00005
Consumer products	170,000,000	0.005–0.03	0.005–0.013
Misc. envir. Sources	~ 25,000,000	0.0006	0.00006
Medical	—	—	0.053

^a1 cSv = 100 crem = 1 rem

^bcirca 1980–1982.

Table 3. Estimated Doses for Space Activities

Scenario	Radiation	Time	Dose, cSv ^{a,b}
Sea level	Various sources	1 yr	~ 0.4
Low Earth orbit	H ⁺	90 d	~ 11
Med. Earth orbit	H ⁺ , e ⁻ , HZE	90 d	~ 7
High Earth orbit	e ⁻ , HZE	90 d	~ 7
Geo. Earth orbit	e ⁻ , HZE	15 d	~ 8
Lunar mission	H ⁺ , e ⁻ , HZE	88 d	~ 7
Mars mission	H ⁺ , e ⁻ , HZE	3 yr	~ 100

^a1 cSv = 100 crem = 1 rem

^bDoes not include any dose from solar events

1 in 2 for males and about 1 in 3 for females (16). Typically, the risk factor in adults for fatal cancer from radiative exposure is about $4 \times 10^{-2} \text{ Sv}^{-1}$ (7, p. 137). Therefore, the recommended upper limits correspond to a fractional increase of a few percent in lifetime risk of fatal cancer.

Doses and Dose Equivalents for Orbital and Interplanetary Missions.

Two important aspects of the assessment of risks from radiation in space remain to be considered: the exposure levels from radiative fields in space and the corresponding risks from those exposures. To address the first issue, measured values of D_{eq} for previous space flights are plotted in Fig. 5 (17). Except for the Apollo missions, these represent relatively short orbital missions compared with either the international space-station or interplanetary missions. Except for the geosynchronous Earth orbit, all of the scenarios result in dose equivalents in the neighborhood of roughly 0.1 cSv per day.

Estimates of True Risk and Associated Uncertainties. As the radiative fields in space environments have continued to be characterized with increasing certainty during the last two decades, we have been able to proceed with estimates of the dose equivalents with increased confidence in the physics. However, the risk estimates continue to exceed established limits, and the errors of such analyses still remain at unacceptable levels for two major reasons:

1. Secondary particles: Although the primary spectra for galactic cosmic rays have been established with reasonable certainty, the variation in the radiative quality within a spacecraft remain both substantial and uncertain because of complex secondary reactions that produce neutrons and secondary charged particles.
2. Biological uncertainty in true risk: Dose equivalent may not represent the true risk of cancer and other diseases in a space environment. Because the particles responsible for radiative damage are more energetic and have greater mass than those encountered at Earth's surface, our experience with this type of radiative environment is severely limited.

Secondary Particles. Figures 6 and 7 present the frequency distribution $y^2 f(y)$ per logarithmic interval of lineal energy y for cobalt-60 gamma rays (18),

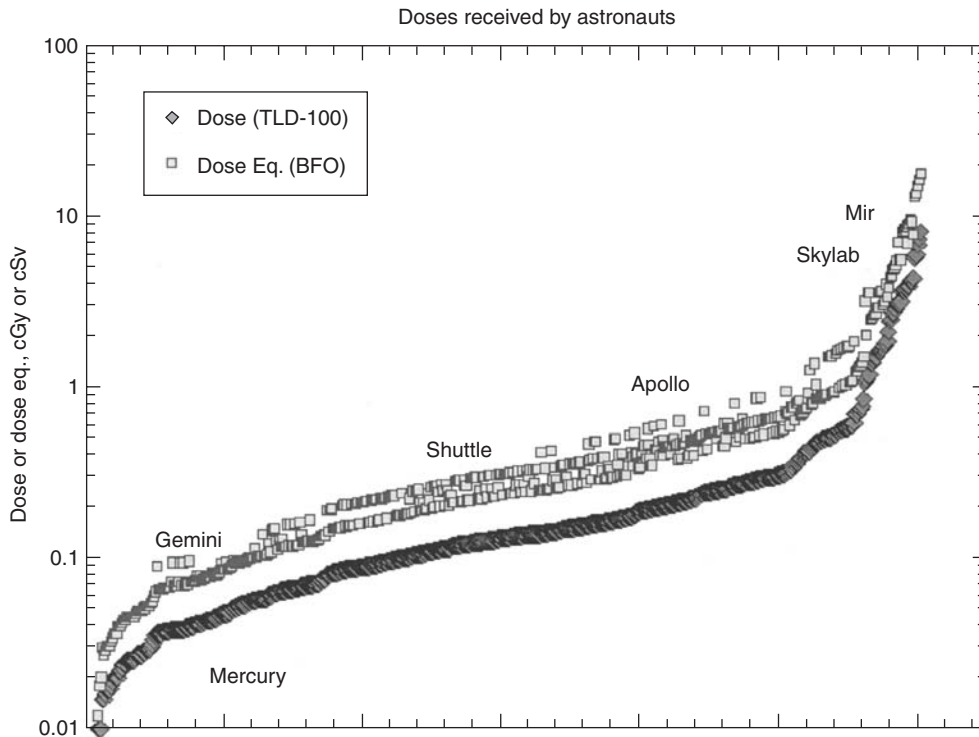


Figure 5. The dose in centigrays and the dose equivalent for blood-forming organs (BFO) in centisieverts for individuals in space (courtesy of NASA Johnson Space Center).

250-MeV protons, 250-MeV/nucleon heavier ions, and neutrons at several energies (18). In the early era of research on risks from radiation in deep space, it was thought that the high-energy heavier ions such as iron were the most biologically damaging and, therefore, the most significant despite their lower abundance. Although protons and helium ions are about a thousand times more abundant, they are sparsely ionizing, that is, have low LET. They were expected to have a biological effect equivalent to that of X or gamma rays. The higher lineal energies from heavier ions of the same energy per nucleon result from a higher ionization density along the tracks. The early conclusion of some researchers was that the increased ionization density or higher LETs of these heavier particles would result in their being biologically more damaging despite their lower abundance. Typical distributions of secondary particles for 800-MeV protons traversing silicon were calculated by J.F. Dicello et al. and others (19). The group of Miller et al. has been measuring the distributions of primary and secondary charged particles produced by energetic heavy ion beams. Although the primary energetic protons are themselves close to minimally ionizing, they produce an abundance of secondary particles, as they traverse the spacecraft or its occupants. These secondaries include an abundance of lower energy protons, neutrons, and ion recoils, and many can be at least as highly ionizing as primary galactic cosmic rays and biologically as effective as primary HZE particles.

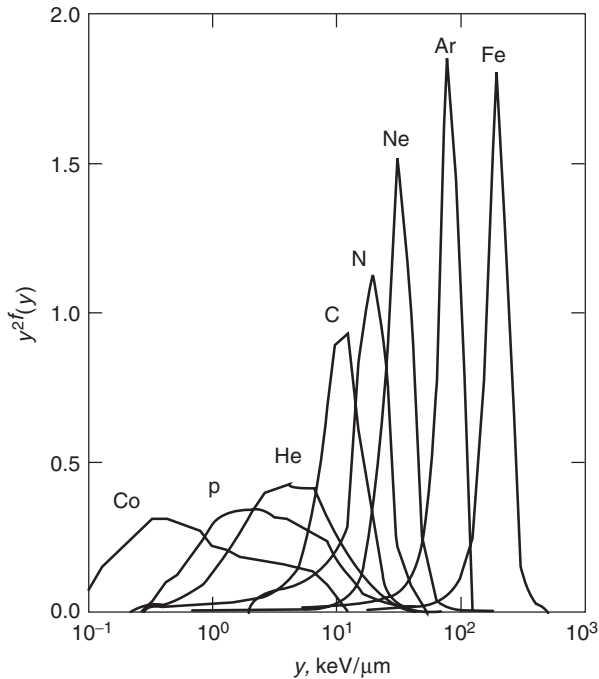


Figure 6. The probability per logarithmic interval of lineal energy $y^2f(y)$ as a function of lineal energy y for gamma rays from cobalt-60, protons of about 200 MeV initial energy, and heavier ions of about 600 MeV per nucleon (courtesy of *IEEE Trans. Nucl. Sci.*).

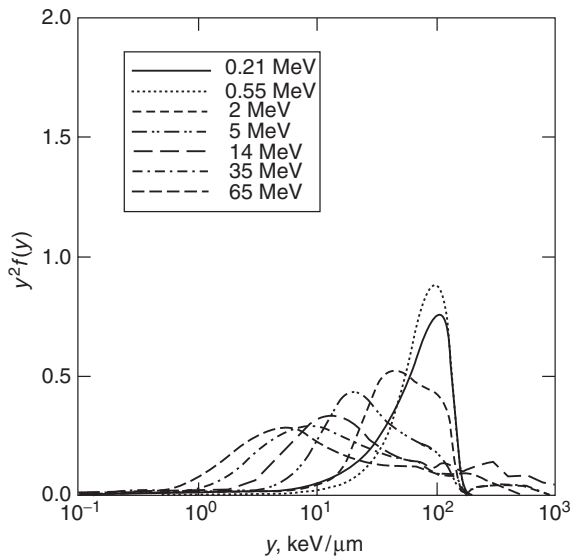


Figure 7. The probability per logarithmic interval of lineal energy $y^2f(y)$, as a function of lineal energy y from external neutron beams of various specified mean energies (courtesy of *IEEE Trans. Nucl. Sci.*).

Based upon the microdosimetric analysis of the primary and secondary particles from galactic cosmic rays, Dicello (20) proposed that primary protons, along with their secondary charged particles and neutrons, might be at least as effective biologically as heavy ions in producing biological damage. The relative doses and dose equivalents expected as a function of depth are presented in Fig. 8 and show that, as the primary protons enter the spacecraft, they would rapidly produce secondaries and become major contributors to the dose equivalent at typical depths in a space vehicle. A major emphasis of current research is to choose materials that can minimize the number of primary HZE particles and also the substantial number of secondaries, particularly secondary neutrons. Despite the gradual change in the premise underlying the research and design, there is little biological data to evaluate any hypothesis.

Biological Uncertainty in True Risk. Although the dose equivalent can be calculated with a reasonable degree of confidence, it remains a major issue as

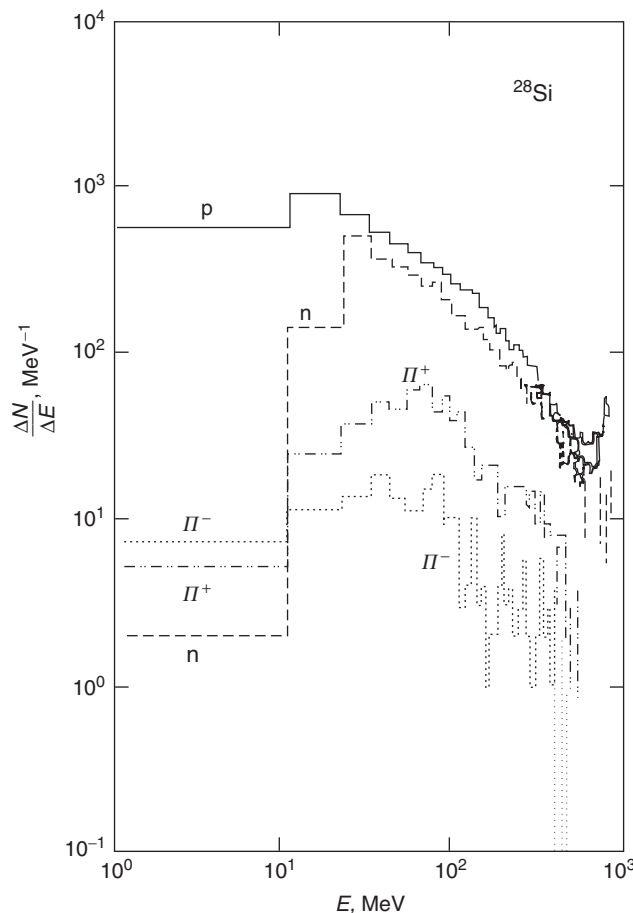


Figure 8. Number of secondary protons, neutrons, and pions per unit energy interval arising from nuclear cascade processes of 800-MeV protons in silicon (courtesy of *Nucl. Methods Instrum.*).

to how accurately the quality factors used to obtain dose equivalent from the physical dose truly represent the actual biological RBEs for cancer or other major diseases. The change in the quality factor as a function of LET remains somewhat arbitrary in radiative protection. We will adopt the relationship described in NCRP 132 (7) and ICRP 60 (13) as the defined relationship for regulatory purposes; however, note that this relationship may not necessarily provide either the correct functional dependency with dose or LET or the correct absolute values within acceptable limits of uncertainty (17). Historically, the relationship between the quality factor and the LET was based initially on animal studies at high doses and cellular end points, primarily *in vitro* survival. Although NASA and other agencies have carried out a life-sciences research program to address these specific issues, to date, only one comprehensive study of carcinogenesis for HZE particles, that of Alpen et al. (21) and Fry et al. (22), has been completed. Fry et al. estimated RBEs for the different particle types at the lower doses of interest for space applications by taking the ratio of the estimated initial slopes of the dose response curves. This group studied the prevalence of tumors that arise in the Harderian gland in the ocular system of mice as a function of particle type and energy. Some of the results are presented in Fig. 9. Their analysis of the results produced RBEs of 27 for iron ions and 4 for helium ions. More recently, Dicello et al. (24) have been measuring the lifetime incidence of breast cancers that arise in a rat model from photons, protons, and iron ions. A large fraction of the animals is still alive, but preliminary results suggest RBEs somewhat lower than those observed by Alpen's group for the Harderian gland.

Countermeasures

Cucinotta et al. (17) attempted to evaluate how well the relationship proposed by ICRP and NCRP conforms to existing data and how much uncertainty still exists

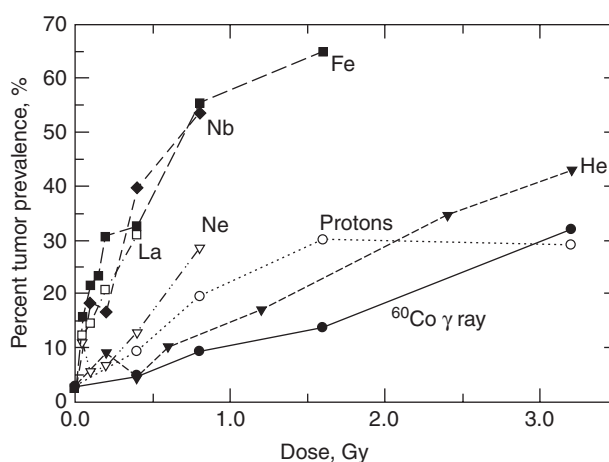


Figure 9. Tumor prevalence vs. dose for selected particles (courtesy of Advanced Space Research).

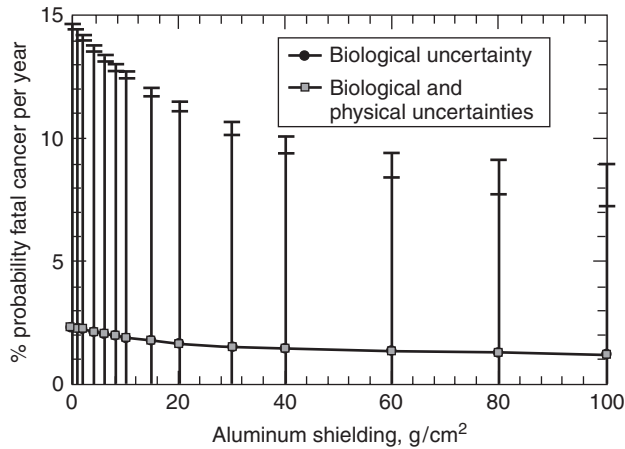


Figure 10. Probability of fatal, cancer incidence as a function of Al shielding thickness (courtesy of Radiation Research).

in proposed scenarios. The result of one error analysis as a function of depth in aluminum is presented in Fig. 10. The range of the vertical bars represents the range of probabilities that cancers will arise from exposures in space that remain likely because of the uncertainties of the measurements and calculations. There is one major conclusion that might be drawn for this study. The uncertainties in the existing data suggest that about 80% of the risk uncertainties result from uncertainties in the quality factors, that is, uncertainties in the relative effectiveness of the different radiations in inducing fatal cancers or other serious diseases. This translates to uncertainty in *in vivo* biological responses, as it relates to human diseases, which further translates into risk factors for potential design criteria, as illustrated in Fig. 11. Cucinotta et al. (17) conclude that the range in values likely for the probability for cancer is as much as a factor of five or more from the mean in specific regions (17,23).

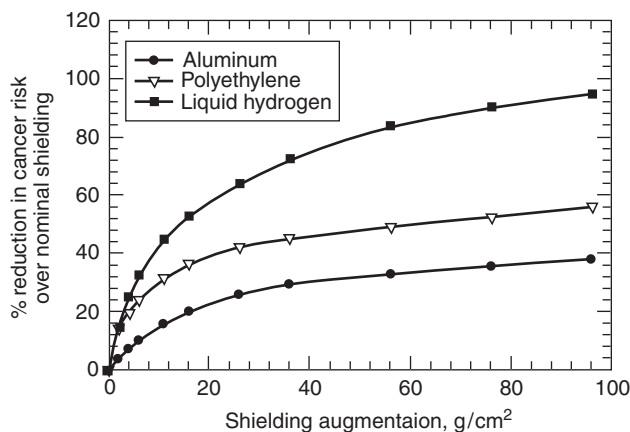


Figure 11. Reduction in cancer risk over nominal shielding as a function of shielding augmentation (courtesy of Radiation Measurements).

Studies are in progress to determine optimal materials and combinations of materials that would reduce the dose equivalents within a spacecraft. Because of the potentially high levels of secondary neutrons, composite designs that include hydrogen-rich materials look particularly promising. Uncertainties in calculating the expected reduction in risk resulting from shielding and the material used for shielding are being reduced through current research to determine better the risks of carcinogenesis from the types of particles present in space (24). At the same time, research to examine pharmaceutical interventions that would reduce the consequences as well as the exposure are underway. Huso et al. (25) have preliminary data to suggest that Tamoxifen[®] administered during the promotion and progressive stages of cancer can reduce the incidence of mammary cancers in rats. This approach has the appeal of being administered only after an exposure (26). Kennedy et al. (26) and Lupton (27) are examining less toxic dietary supplements to reduce cancer, and Vazquez (28) is examining the effectiveness of drugs to reduce damage to the central nervous system.

Summary

The risk of cancer and other diseases arising from radiative exposures in space may be the major biomedical risk for missions of long duration because of the level of risk and the uncertainties of reducing such risks to acceptable levels. These risks arise primarily from exposures to galactic cosmic rays, particles from solar events, and particles trapped in Earth's magnetic fields, in decreasing order of potential significance. Considerable progress has been achieved in characterizing the physical characteristics of the primary radiative fields that surround Earth and are also present in space. Less progress has been made in determining the biological risk of cancer per unit exposure for the different types of radiation and for the probable time periods. Our best estimates suggest unacceptable levels of risk, although what constitutes an acceptable risk for missions beyond Earth's magnetosphere has yet to be defined. Research continues to make headway toward resolving the major issues, and few experts in the field doubt that safe interplanetary missions will be possible in the near future.

Note

This work was supported in part by NASA Grant NCC9-58, National Space Biomedical Research Institute.

Dr. John F. Dicello, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Division of Radiation Oncology, The Harry and Jeanette Weinberg Bldg., 401 N. Broadway, Suite 1400, Baltimore, MD 21231-2410.

BIBLIOGRAPHY

1. NASA Critical Path Roadmap, 11-17-2000. <http://criticalpath.jsc.nasa.gov>.
2. Setlow, R. *Radiation Hazards to Crews of Interplanetary Missions: Biological Issues and Research Strategies*. National Academy Press, Washington, DC, 1996.

3. Osborn, M.J. *A Strategy for Research in Space Biology and Medicine in the New Century*. National Academy Press, Washington, DC, 1998.
4. Van Allen, J.A. The First Public Lecture on the Discovery of the Geomagnetically Trapped Radiation. State University of Iowa, Report 60-13, 1960.
5. Atwell, W. Personal communication as referenced in NCRP Report 132, 1999.
6. Wilson, J.W., L.W. Townsend, W. Schimmerling, G.S. Khandelwal, F. Khan, E. Nealy, F.A. Cucinotta, L.C. Simonsen, J. Shinn, and J.W. Norbury. *Transport Methods and Interactions for Space Radiations*. NASA Reference Publication 1257, Scientific and Technical Information Program, 1991.
7. National Council on Radiation Protection and Measurements (ed.). *Radiation Protection Guidance for Activities in Low-Earth Orbit*, NCRP Report 132. National Council on Radiation Protection and Measurements, Bethesda, MD, 2000.
8. Ebert, H.G. (ed.). *Microdosimetry*. European Communities, Brussels, ICRU Publications, Belgium, 1968.
9. ICRU Report 36. *Microdosimetry*. ICRU Publications, Bethesda, MD, 1983.
10. Dicello, J.F. Practical implications of microdosimetry. In U. Linz (ed.), *Ion Beams in Tumor Therapy*. Chapman & Hall, New York, 1995, pp. 35–44.
11. Rossi, H.H., and M. Zaider. *Microdosimetry and Its Applications*. Springer-Verlag, Berlin, 1996.
12. *Guidance on Radiation Received in Space Activities*. National Council on Radiation Protection and Measurements Report 98, Bethesda, MD, 1989.
13. *1990 Recommendations of the International Commission on Radiological Protection*. ICRP Publication 60. Pergamon, Oxford, England, 1991.
14. *Implementation of the Principle of As Low As Reasonably Achievable*, NCRP Report 107. NCRP, Bethesda, MD, 1990.
15. *Health Effects of Exposure to Low Levels of Ionizing Radiation: BEIR V*. National Academy Press, Washington, DC, 1990.
16. *Cancer Facts & Figures 2001*. American Cancer Society, New York, 2001.
17. Cucinotta, F.A., W. Schimmerling, J.W. Wilson, L.E. Peterson, G.D. Badhwar, P.B. Saganti, and J.F. Dicello. *Space Radiation Cancer Risk Projections for Exploration Missions: Uncertainty Reduction and Mitigation*. NASA TP-2002-210777 NASA Langley Research Center, 2001.
18. Dicello, J.F., M. Wasiolek, and M. Zaider. Measured microdosimetric spectra of energetic ion beams of Fe, A, Ne, and C: Limitations of LET distributions and quality factor in radiation effects and space research. *IEEE Trans. Nucl. Sci.* 38: 1203–1209 (1991).
19. Dicello, J.F., M.E. Schillaci, and L. Liu. Cross sections for pion, neutron, proton, and heavy-ion production from 800-MeV protons incident upon aluminum and silicon. *Nucl. Instrum. Methods B45*: 135–138 (1990).
20. Dicello, J.F. HZE cosmic rays in space. Is it possible that they are not the major radiation hazard? *Radiat. Prot. Dosimetry* 44: 253–258 (1992).
21. Alpen, E.L., P. Powers-Risius, S.B. Curtis, R. DeGuzman, and R.J.M. Fry. Fluence-based relative biological effectiveness for charged particle carcinogenesis in mouse harderian gland. *Adv. Space Res.* 14 (10): 573–581 (1994).
22. Fry, R.J.M., P. Powers-Risius, E.L. Alpen, E.J. Ainsworth. High-LET radiation carcinogenesis. *Radiat. Res.* 104: S-188–S-195 (1985).
23. Cucinotta, F.A., J.W. Wilson, J.R. Williams, and J.F. Dicello. Analysis of MIR-18 results for physical and biological dosimetry: Radiation shielding effectiveness in LEO. *Radiat Meas.* 32 (3): 181–191 (June 2000).
24. Dicello, J.F. Radiation effects: Core Project. In J.F. Dicello, F.A. Cucinotta, D.S. Gridley, S.P. Howard, D.L. Huso, G.R. Novak, P.A. Overbeek, S. Piantadosi, and J.D. Strandberg (eds), *Life Sciences Tasks and Bibliography for FY 2000*. <http://peer.nasaprs.com>, March 1, 2002.

25. Huso, D.L., and J.F. Dicello. Chemoprevention of radiation-induced neoplasms. Life Sciences Tasks and Bibliography for FY 2000, <http://peer.nasaprs.com>., March 1, 2002.
26. Kennedy, A.R., W. Troll, and J.B. Little. Role of free radicals in the initiation and promotion of radiation transformation in vitro. *Carcinogenesis* 5: 1213–1218 (1984); Kennedy, A. Countermeasures for Space Radiation Biological Effects. <http://www.nsbri.org/Research/2001-2003>, March 1, 2002.
27. Lupton, J. Nutritional Countermeasures to Radiation Exposure. <http://www.nsbri.org/Research/2001-2003>, March 1, 2002.
28. Vazquez, M. CNS Damage and Countermeasures. <http://www.nsbri.org/Research/2001-2003>, March 1, 2002.

J.F. DICELLO
Johns Hopkins University School of Medicine
Baltimore, Maryland

F.A. CUCINOTTA
NASA Johnson Space Center
Houston, Texas

SPACE RESOURCES, OCCURRENCE AND USES

Space resources consist of all of the useful materials, energy sources, and energy found in space. For practical purposes, this article focuses on those resources that may be accessible within the next 25 years, and for which large-scale demand is plausible. Therefore, the principal focus of this review is the bodies of nearby space: Earth's Moon, the planets Mercury and Mars, the two small Martian moons, Phobos and Deimos, and the near-Earth population of asteroids and comets.

This article briefly surveys the significance and promise of space resources, including scientific and technical issues and the legal and regulatory regime surrounding their economic use. Current treaty obligations are reviewed, and some suggestions made regarding legislation and treaty language that would encourage the growth of this enormous new arena of economic activity, making vast new resources available to all.

Why Use Space Resources?

The cost of access to space is presently very high. To lift one kilogram from Earth's surface into low Earth orbit costs about \$10,000: a tonne of water aboard a space station represents an investment of nearly \$10,000,000 in launch costs. Soft-landing one tonne of payload on the Moon costs about \$100,000,000. A gallon of gasoline on the Moon would cost \$400,001—\$1 to purchase the gas on Earth and \$400,000 to deliver it to the Moon.

Such a pricing system seems intuitively absurd: is there no better way to make intrinsically cheap, common, and highly desired materials available in

space at affordable prices? How could we do better than launching them from Earth? Certainly any source of water, life-support materials, or propellants already present on the Moon or Mars would be enormously attractive. Even if it cost us \$100 to extract or manufacture a liter of propellants on the surface of another planet, that amount would represent a 99.9% cost reduction, compared to transporting those propellants from Earth. We could then carry out the same mission for 0.1% of the propellant cost of launching the propellants from Earth—or move 1000 times as much payload for the same cost. To exercise this option, we require a scientifically, technically, and economically sound extraterrestrial source of propellants. This principle also applies to other commodities that are intrinsically cheap and in great demand in space, such as metal construction materials for building large space structures.

In recent years a number of suggestions have been offered regarding economic exploitation of the energy and material resources in nearby space. Among the resources proposed for exploitation are

1. solar energy, from space-based photovoltaic cells (Solar-Power Satellites or lunar power stations), via microwave transmission of power to antennae on Earth, to provide clean, cheap energy in vast amounts;
2. oxygen extracted from lunar or asteroidal rocks and minerals for use in life-support systems and as the oxidizer for rocket engines departing from the Moon;
3. the isotope helium-3, a constituent of the atmospheres of Uranus and the other giant planets and a very minor constituent of lunar regolith, for use as clean fusion fuel on Earth;
4. water from near-Earth asteroids, Martian permafrost, or lunar polar ice deposits for use in life support and rocket propellants;
5. metals, especially ferrous-metal alloys, for use in space construction; and
6. precious and strategic metals for importation to Earth.

As these examples suggest, the most important single motivation for space resource use is the recognition that large-scale future activities in near-Earth space may be made far less expensive by using materials that are already in space and that therefore need not be launched at great expense out of Earth's gravity well. But this is not the only motivation. A second important benefit of using space-derived energy and resources is the enormous positive environmental impact of off-loading energy-related mining, drilling, and shipping and large-scale energy production from Earth's surface. Return of valuable commodities to Earth merits third priority.

Many material resources found in space, especially on the Moon and Mars, would be used locally (i.e., within the gravity well of the body from which the resource was mined) to support unmanned and manned exploration or to fuel vehicles returning to Earth. Because of the substantial gravity fields of Mars and the Moon and the consequent high energy cost of landing and takeoff, as well as the low quality of their proven mineral resources, profitable export of materials from such massive bodies is far less feasible than local use: neither of these two bodies is well suited to export materials for use in other locations. For these

logistical reasons and because of their great resource richness and diversity, near-Earth asteroids have emerged as the most attractive nearby sources of space-derived materials for export (especially, for return to near-Earth orbits).

The realization that these space resources may have enormous economic impact has stimulated considerable technical and economic interest. Two technical volumes on space resources have appeared in recent years, *Space Resources* (1) and *Resources of Near-Earth Space* (2). A popular overview of space resources based on these technical studies appears in *Mining the Sky* (3). A series of conference volumes containing papers on space resources, based on the proceedings of the biennial Princeton High Frontier Conferences and entitled *Space Manufacturing* (volume numbers I to XIV), has been published by the American Institute of Aeronautics and Astronautics. The technical literature on space resources is otherwise very widespread and difficult to research; it often appears in contractor reports, limited-circulation conference proceedings, and government publications. For this reason, using the volumes mentioned as introductions to the literature and as a source of citations to original publications is highly desirable.

The Resources of the Moon. The best single source for information on the physical and chemical properties of the Moon is the *Lunar Sourcebook* (4). A detailed treatment can be found in the article, *The Moon* in this volume. General reviews of and references to processing of lunar materials are found in the basic references cited earlier.

The Moon is in orbit around Earth at a mean distance of 384,400 km from Earth's center. Its orbit has an eccentricity of 0.055 and an inclination of 5.142° relative to the ecliptic that allows excursions of the Moon to as far as 28.7° north or south of the terrestrial equator. The Moon is airless and utterly devoid of liquid water. Its radius is 1738 km and its mass is $1/81.3$ times the mass of Earth. Its surface gravity is 0.165 Earth gravities, and its escape velocity is 2380 m s^{-1} .

The Lunar Rocks. The Moon is largely covered by heavily cratered highlands that are rich in the calcium aluminosilicate mineral, anorthite ($\text{CaAl}_2\text{Si}_2\text{O}_8$), one of the two defining constituents of terrestrial plagioclase feldspar. The Moon is so severely depleted in alkali metals that the complement of sodium plagioclase (albite; $\text{NaAlSi}_3\text{O}_8$) normally found on Earth is reduced to a minor component on the Moon. The most common highland rock type is anorthosite, a rock dominated by anorthite. The second most abundant mineral in these rocks is generally a low-magnesium pyroxene, and olivine is also present. These rocks, the ferroan anorthosites, are among the most ancient rocks on the Moon and date back at least 4.4 Ga (billion years before the present).

The second most abundant highland rocks are magnesium-rich and compositionally diverse. They have plagioclase contents that range to nearly zero in dunite, which is nearly pure olivine rock. Olivine-plagioclase rocks are called norites, and pyroxene and Ca-poor plagioclase rocks are classified as gabbros or norites. Many of the Mg-suite highland rocks also date to before 4.4 Ga.

An unusual lunar highland rock type that contains enhanced abundances of potassium (K), the rare-earth elements (REE), and phosphorus (P), found only as small chips in the regolith, has been named KREEP. In major-element chemistry, KREEP is a basalt. As on Earth, the rare earths concentrate strongly in phosphate minerals. Gamma-ray spectroscopy conducted from lunar orbit during the Apollo program tentatively identified the KREEP source region on the edge of the

Mare Imbrium basin, based on the enhanced gamma emission from high concentrations of K, U, and Th. These data have recently been confirmed and refined by a gamma-ray spectrometer on the Lunar Prospector mission.

Chemical analyses of representative lunar igneous rocks show that most have silica abundances that are near the low end of the terrestrial range; they usually have less than 50% by weight of SiO_2 , within the range of terrestrial basalts. Materials from the lunar lowlands, which cover about 25% of the near side of the Moon, are typically basaltic in composition.

However, lunar basalts differ from terrestrial basalts in that the lunar rocks have alkali metal abundances that are several times lower and titanium abundances that are several times higher than the usual terrestrial examples. Because the very high titanium abundance is accounted for by large amounts of ilmenite (FeTiO_3), the iron abundance is also abnormally elevated in lunar basalts relative to terrestrial basalts.

We find a range a range of textures, from the most fine-grained lunar basalts, which cooled rapidly upon extrusion as lava flows onto the surface of the Moon, to coarser grained basalts that crystallized more slowly and completely in a more protected (often intrusive) environment beneath the surface, forming volcanic sills or dikes or the bottoms of thick lava flows. Extruded basalts often show abundant vesicles, essentially bubbles inflated by magmatic gases such as hydrogen or CO, which unfortunately have long since diffused away.

Liquids of lunar basalt composition have very low viscosities and therefore are poorly suited for building tall volcanic structures about their magma vents. They are, however, ideal for making lava tubes, which may later drain and collapse to make features that look like the lunar rilles. These basaltic liquids have densities of only about 3.0, compared to about 3.3 for the solidified basalt. The observed bulk density of the Moon is only 3.34.

Lunar Minerals of Interest as Resources. Only seven minerals are ever found in abundances greater than about 1% in lunar rocks (Table 1). These include pyroxene (Ca,Fe,MgSiO_3); calcic plagioclase ($\text{Ca,Na(Al,Si)}_4\text{O}_8$, close to anorthite composition; ilmenite (FeTiO_3); olivine ($\text{Mg,Fe}_2\text{SiO}_4$); pyroxferroite $\text{CaFe}_6(\text{SiO}_3)_7$; and two polymorphs of silica (SiO_2), cristobalite and tridymite.

Pyroxene compositions in lunar rocks trespass into regions of the pyroxene quadrilateral not previously populated by other solar system materials. Some lunar pyroxenes are fairly standard augites, but others (often the outer layers of normal augite crystals) range far into the ferrosilite corner of the quadrilateral, perilously close to pure FeSiO_3 , which is thermodynamically unstable with respect to its component oxides. This new material crystallizes in the triclinic system, unlike monoclinic augite, and hence is given a new mineral name, pyroxferroite. Pigeonite, which is low in calcium, is also present in smaller amounts.

Feldspars are present as an unusually calcium-rich plagioclase (from about 60% up to 99% anorthite) and a potassium feldspar. The K-spar, which has a sanidine structure, is crystallized in very small amounts out of the residual melt in rapidly crystallized basalts.

Given the low silica content of most lunar rocks, the presence of small amounts of olivine is hardly surprising. Olivine is well approximated by a solid solution of Mg_2SiO_4 (forsterite; fo) and Fe_2SiO_4 (fayalite; fa). Most lunar olivine is fairly iron-rich ($\text{fa}_{20}\text{-fa}_{50}$), but virtually pure fayalite is sometimes observed.

Table 1. Selected Native Lunar Minerals

<i>Metals</i>		
Kamacite	Fe,Ni (<6% Ni)	a ^a
Taenite	Fe,Ni (>6% Ni)	a
<i>Sulfides</i>		
Troilite	FeS	a
<i>Oxides</i>		
Armcolite	FeMgTi ₂ O ₅	a
Perovskite	CaTiO ₃	a
Spinel S.S. ^b		a
Spinel	MgAl ₂ O ₄	
Hercynite	(Fe,Mg)Al ₂ O ₄	
Chromite	FeCr ₂ O ₄	
Magnesiochromite	MgCr ₂ O ₄	
Ulvospinel	Fe ₂ SiO ₄	
Cristobalite	SiO ₂	m ^a
Tridymite	SiO ₂	m
Rutile	TiO ₂	a
Baddeleyite	ZrO ₂	a
Ilmenite	FeTiO ₃	M
<i>Oxysalts</i>		
Fluorapatite	Ca ₅ (PO ₄) ₃ F	a
Chlorapatite	Ca ₅ (PO ₄) ₃ Cl	a
Whitlockite	Ca ₃ (PO ₄) ₂	a
<i>Silicates</i>		
Olivine S.S. ^b	(Mg,Fe) ₂ SiO ₄	m
Fayalite	Fe ₂ SiO ₄	
Forsterite	Mg ₂ SiO ₄	
Pyroxene S.S. ^b		M ^a
Orthopyroxene	(Mg,Fe)SiO ₃	
Enstatite	MgSiO ₃	
Ferrosilite	FeSiO ₃	
Clinopyroxene	(Ca,Mg,Fe)SiO ₃	
Wollastonite	CaSiO ₃	
Feldspar S.S. ^b		M
Plagioclase		
Anorthite	CaAl ₂ Si ₂ O ₈	
Albite	NaAlSi ₃ O ₈	
K-spar		a
Orthoclase	KAlSi ₃ O ₈	
Sanidine	KAlSi ₃ O ₈	
Pyroxferroite	CaFe ₆ (SiO ₃) ₇	m
Zircon	ZrSiO ₄	a

^aM: major mineral (>10% in some rocks); m: minor mineral (1 to 10% in some rocks); a: accessory mineral (always <1%).

^bS.S. Solid solution.

Fayalite, unlike pure ferrosilite, is thermodynamically stable. Many lunar basalts are close to or above silica saturation and often contain a few percent of cristobalite or tridymite or even a trace of quartz.

Ilmenite occurs in basalts at abundances from a few percent to more than 20%. There is an important difference in oxidation state between terrestrial ilmenite, which is a solid solution of FeTiO_3 and Fe_2O_3 , rich in ferric iron, and lunar ilmenite. Ferric iron is absent from lunar igneous rocks, and in fact small amounts of metallic iron are often found in lunar basalts. The ferric mineral FeOOH , ubiquitous in trace amounts in lunar samples, is an alteration product produced by the attack of terrestrial atmospheric water vapor on lunar lawrencite (FeCl_2).

Metallic iron in lunar rocks, generally found in association with troilite, is nearly pure iron and contains less than 1% nickel. The molar ratio of metal to troilite always lies close to the eutectic composition for an Fe–FeS melt. On rare occasions, tiny traces of metallic copper are also found in association with the metal and troilite in basalts.

By contrast, metallic iron found in the regolith and in the shock-lithified microbreccia contains up to 30% Ni and 1% Co. The metal in the basalt seems to have been made by reduction of FeO during melting, whereas the regolith metal is clearly dominated by asteroidal debris. This view is reinforced by the common occurrences of traces of cohenite (Fe_3C) and schreibersite $[(\text{Fe},\text{Ni})_3\text{P}]$, both common accessories of meteoritic metals, in association with Ni-bearing metal in the regolith and breccias.

Several interesting oxide minerals besides ilmenite are also found on the Moon (Table 1). Rare spinel, nearly stoichiometric MgAl_2O_4 , has been found in breccias, chromite FeCr_2O_4 has been found within regolith nickel-iron particles of apparent asteroidal origin, and ulvospinel, nearly stoichiometric Fe_2TiO_4 , has been found in trace amounts exsolved from or replacing ilmenite. Other titanates include armalcolite $[(\text{Fe},\text{Mg})\text{Ti}_2\text{O}_5]$, usually found within ilmenite grains, and perovskite CaTiO_3 that contains high concentrations of rare-earth elements, in the late-crystallizing component of the coarser basalts. Rutile TiO_2 and both baddeleyite (ZrO_2) and zircon (ZrSiO_4) are also found in tiny quantities.

Few elements serve as important markers of the oxidation state of lunar material. We have seen lunar basalts lie close to the Fe–FeO buffer. Sulfur is fully reduced to sulfide (indeed, almost exclusively as troilite), and carbon is found as meteoritic carbide. Phosphorus, which is found as coexisting phosphide and phosphate in some meteorites, is found in accessory amounts in lunar basalts as apatite $[\text{Ca}_5(\text{PO}_4)_3\text{X}]$. Fluorapatite ($\text{X} = \text{F}$) seems to be more common, but chlorapatite ($\text{X} = \text{Cl}$) has also been reported. Hydroxylapatite ($\text{X} = \text{OH}$) has not been found. Small amounts of an amphibole that contains fluoride instead of hydroxyl have also been reported. Whitlockite $[\text{Mg}_3(\text{PO}_4)_2]$ that has very high concentrations of rare earths (to about 10%) has also been found as a component of KREEP.

The Lunar Regolith. The lunar regolith, the crushed debris layer that covers the surface, is extremely complex. It is dominated by the products of violent cometary and asteroidal impacts on lunar igneous rocks and on the regolith itself. Rocks of all sizes, from house-sized boulders down to tiny rock chips and dust that contains only one or a few mineral grains, are mixed together with tiny glass droplets, similar to small chondrules, and with dust composed of very finely

crushed grains. Impacts in the regolith have produced rocks, called microbreccias, that are essentially shock-lithified samples of the lunar regolith. Eclectic mixtures of smaller particles are often also found welded together by melt glass from later impacts. These welded, highly heterogeneous lumps are called agglutinates. In some places, the lunar regolith has been reprocessed so thoroughly that it has been mostly converted into agglutinates. Agglutinate-rich regolith samples are said to be *mature*.

Polar Volatiles on the Moon and Mercury. Lunar polar ice was apparently first envisioned as a resource by American rocket pioneer Robert H. Goddard in his student notebooks that date from 1908–1910. Harrison Brown and co-workers at Cal Tech in the early 1950s presented a simple quantitative argument for the preservation of ice from water-bearing impactors by recondensation in the lunar polar regions. These calculations were extended in the 1960s by James A. Arnold at UCSD, whose calculations stimulated considerable interest among both theorists and builders of spacecraft instrumentation.

The physical search for polar volatiles on the Moon and Mercury sounds at first like an exercise in futility. The problem is not just the expected very low abundance of volatiles in both bodies, but the logic of detection: the only places on the Moon and Mercury that are cold enough to permit trapping and long-term retention of volatiles (roughly 100 K for water ice) are permanently shadowed regions, such as crater bottoms, very close to the poles. Because the ice must be permanently shadowed to survive, it is always in the dark and cannot be photographed. Further, small, very cold regions on a generally very hot planet are difficult to detect in the infrared, where the long wavelengths of thermal infrared radiation degrade our spatial resolution and high fluxes from hot spots completely swamp the tiny fluxes from cold regions. The intensity of emitted thermal radiation from 600 K areas is greater than that from 100 K areas by a factor of $(T_{\text{hot}}/T_{\text{cold}})^4$, or about 1300:1. Temperatures below 110 K are necessary to keep the evaporation rate low enough to preserve water ice for billions of years.

The first probe of the polar regions of Mercury was an ingenious experiment carried out in 1991 by Martin Slade of the Jet Propulsion Laboratory and Bryan Butler and Duane Muhleman of Cal Tech. They used the 70-meter radar transmitter at Goldstone in the Mojave Desert to illuminate Mercury with monochromatic (single-frequency) radar pulses at a wavelength of 3.5 cm. Slade and his co-workers transmitted a right circularly polarized (RCP) radar signal that would reflect from a flat mirror-like surface with left circular polarization (LCP). The center of the disk of Mercury is essentially a flat, though rough, mirror, so there is a large intensity spike in the returned signal that has minimum time delay, zero Doppler shift (after allowing for the relative motion of the transmitter and the center of Mercury), and strong left-hand polarization. Extremely complex scattering from a very rough surface tends to depolarize the returned signal, which means a relative increase in the RCP component. When the Goldstone/VLA data were analyzed, a bright RCP component was found at zero Doppler shift and maximum range, which requires that it originate at a pole. Fortunately, Mercury's orbit is significantly tilted with respect to the plane of the ecliptic and affords observers on Earth frequent opportunities to see either pole. At the time of observation, the North Pole of Mercury was slightly tipped toward the observers and the South Pole was invisible, so the feature clearly must be associated with the

north polar regions. Because the radar scattering properties observed for this feature are, so far as is known, unique to ice-covered surfaces, the strong implication is that Mercury has polar ice. Since these original observations, studies at different Earth–Mercury geometries have revealed that a similar feature is also present near the South geographic Pole. That area of anomalous reflection appears to be confined mostly to the floor of the large Chao Meng-Fu impact crater, which is centered at 87.5° S latitude.

Polar ice need not be exposed on the surface. Tens of centimeters of dry, porous dirt could overlie the ice without degrading the distinctive radar reflection signature of the ice. The thickness of the ice layer is poorly constrained by the observations. A few tens of centimeters or a few meters of ice would suffice to explain the observations, but a few kilometers of ice would be possible. Further, other materials besides water ice might possibly look the same.

Thermal modeling of the polar regions by David Paige and co-workers at UCLA shows that flat, unshadowed regions near the poles can be as cold as 167 K and that the observed cratering of the polar regions favors the existence of small, permanently shadowed enclaves inside craters that have large depth to diameter ratios. The craters that exhibit the highest depth to diameter ratios are simple bowl-shaped craters that have rim-crest diameters of 10 km or less. Unfortunately, because of the geometry of the Mariner 10 – Mercury encounters, our photographic (visible light) coverage of the polar regions has only an extremely narrow range of solar longitudes and leaves nearly half of the polar regions unimaged. The available coverage does, however, suggest areas that may be perpetually colder than 100 K, and even as low as 60 K. The coldest spot inside Saturn's orbit may be on Mercury!

The knowledge that ices can be stable in the Moon's and Mercury's polar regions is fascinating but does not by itself tell us the source of the volatiles that are condensed there. Outgassing from their interiors seems an unlikely source if these planets are made of volatile-poor, high-temperature condensate; however, the mass of ice required to explain the observations may be as low as a few cubic kilometers, an amount so small as to make it impossible to rule out the presence of that much intrinsic water in the planet. Further, cometary impacts during billions of years provide vastly larger fluxes of water onto both bodies than the amount required by the radar data.

Long-period comet impacts can be enormously energetic events. At Mercury's mean orbital speed of about $V_{\text{orb}} = 48 \text{ km s}^{-1}$, a low-inclination retrograde parabolic comet could encounter Mercury at a speed as high as $(1 + \sqrt{2})V_{\text{orb}}$, about 116 km s^{-1} , a kinetic energy density of $6.7 \times 10^{13} \text{ erg g}^{-1}$. Mercury's escape velocity is only 4.3 km s^{-1} (an escape energy of $9 \times 10^{10} \text{ erg g}^{-1}$), so an impacting long-period comet carries enough energy to eject its own mass, plus more than 740 times its mass of Mercury's regolith, from the planet. This is not an efficient method of emplacing water on Mercury. Short-period comets, however, have average encounter speeds that are several times lower and have minimum encounter speeds of the order of 10 km s^{-1} . More important, extinct periodic comets (that have dust-insulated ice cores) and C-type planet-crossing asteroids may occasionally approach at velocities not much greater than escape velocity. Such impacts carry vastly less energy per gram and also have higher mean atomic weights in their fireballs, and hence expand more slowly. Thus water-bearing

asteroids appear to be the most credible source of water and other volatiles on Mercury.

Unfortunately, our knowledge of the velocity distribution of Mercury-crossing asteroids is still very incomplete. Recent Spacewatch discoveries, mentioned later, show a surprising number of 10- to 500-m bodies in extremely Earth-like orbits of low eccentricity and modest inclination. The origin of this newly discovered class of bodies is not understood, and hence the likelihood of finding such a swarm of low-velocity bodies near Mercury cannot be assessed. Note also that, although there are many known Mercury-crossers among the near-Earth asteroid population, they represent an extremely biased set: all of them were discovered in the night sky from Earth and therefore must be in high-eccentricity orbits. Conversely, any low-eccentricity bodies that orbit near Mercury (or Venus) could not be discovered by present asteroid search techniques. Further, because all Mercury-crossing asteroids so far discovered must also cross Venus and Earth, they are strongly depleted in bodies whose inclinations are so low that they make frequent close approaches to these three planets.

One important implication of the discovery of polar ice on Mercury is that it suggested a similar phenomenon on the Moon. If there are massive quantities of polar volatiles on the Moon, both rocket propellants and life-support fluids could be manufactured readily on the lunar surface. Such a local source of propellants would greatly decrease the cost of launch operations from the Moon and possibly provide a source of propellant for export to other locations in the Earth–Moon system, most of which are far more accessible from the lunar surface than from the surface of Earth.

The first evidence of the presence of lunar polar ice deposits came from the Clementine mission in 1994. The Radio Science experiment on this spacecraft took advantage of the observed “anomalous depolarization” of radio signals reflected from an ice surface, a phenomenon discovered in the course of radar studies of the icy Galilean satellites of Jupiter. The Clementine observations showed that when signals from the spacecraft transmitter are bounced at grazing incidence off the lunar poles, they were depolarized in the same way as those reflected from Europa, Ganymede, and Callisto—and the Mercurian poles. Although there is no convincing theoretical reason to link this phenomenon exclusively to water ice, it has not yet been shown that any other natural substance has the same effect. The data suggested that a small fraction of the surface area immediately adjacent to the lunar poles, possibly permanently shadowed crater bottoms, contains ice.

Clementine’s indirect detection of ice set the stage for the much more specific neutron spectroscopy carried out by the Lunar Prospector spacecraft. This experiment is diagnostic for the presence of abundant hydrogen. Identification of the hydrogen-bearing material as water ice is inferential, but ice is the most abundant hydrogen compound in the Universe and has a vapor pressure compatible with the deduced latitudinal distribution of ice on the Moon (and Mercury). To a chemist, the exact molecular speciation of hydrogen is of enormous interest, especially because these hydrogen compounds are almost certainly of cometary or asteroidal origin. But to a chemical engineer, it hardly matters which ices are present: reacting any of them with FeO at high temperatures releases water in abundance. Lunar Prospector found ice deposits

on crater floors near the poles with a total mass probably of the order of a billion tonnes (10^{15} grams). The ice is probably buried at a shallow depth in the regolith as interstitial ice, or permafrost, and partially fills the pore volume of the regolith.

Lunar Resource Exploitation. Many authors have suggested practical uses for lunar materials (Table 2). The simplest such use would be to employ regolith as radiation and micrometeoroid shielding for a lunar base. Using a sufficient power source, chemical processing of the regolith can also be attempted.

The most frequently discussed lunar product is oxygen, liberated by reduction of oxides of iron and possibly other metals (5). This process naturally makes metals, especially iron, available as an ancillary product. Ferrous oxide, an abundant and easily reduced component of many lunar minerals (especially pyroxene $[(\text{Mg},\text{Fe},\text{Ca})\text{SiO}_3]$, olivine, $[(\text{Fe},\text{Mg})_2\text{SiO}_4]$, and ilmenite (FeTiO_3)) can be reduced at elevated temperatures by gaseous reagents such as hydrogen, carbon monoxide, and methane. The oxygen-bearing product, water vapor and/or carbon dioxide, can then be split electrochemically to release oxygen and reconstitute the original reducing agent.

In principle, it would be desirable to heat and process only chemically pure mineral separates. However, beneficiation, the process of separating high-grade

Table 2. **Resource Targets on the Moon**

Product	Resource target	Process
Shielding	Regolith	Mechanical
Oxygen	Ilmenite	CO reduction
		H ₂ reduction
		Methane/HC reduction
		H ₂ SO ₄ dissolution
		Li/Na reduction
		Plasma reduction
	Mare regolith	H ₂ glass reduction
		Magma electrolysis
		Strong base solution
		Carbon reduction
		Vapor pyrolysis
		HF dissolution
	Highland rock	Magma electrolysis
		Strong base solution
		Carbon reduction
Fe/Ni metals	Regolith	Magnetic separation
		Any reduction process
		Magma electrolysis
	Ilmenite	Any reduction process
Water	Regolith	Solar wind H release
	Polar ice deposits	Melting, distillation
Refractories	Regolith	Electrolytic residue
	Ilmenite	Any reduction process
Helium-3	Ilmenite	Heating, gas fractionation

mineral fractions, is difficult on the Moon for a number of reasons. First, “static cling” makes powders extremely sticky in the lunar vacuum and inhibits electrostatic or magnetic separation. Second, ilmenite-bearing rock chips have other minerals firmly bonded to the ilmenite, preventing clean separation. Third, mature regolith has been extensively processed into agglutinates, which weld together grains of disparate compositions. Fourth, the glassy component of the agglutinate is a mutual solution of many minerals that have been rendered inseparable by melting.

If ferrous metals are sought, ilmenite reduction is a logical technique: the process (ignoring dross and unreacted ilmenite) produces oxygen, iron, and the refractory oxide, rutile (TiO_2). Passing the iron through the gaseous carbonyl (Mond) process produces high-purity iron that is so free of defects that it resists corrosion as well as stainless steel.

Other processes for extracting oxygen, including high-temperature thermal decomposition of oxides, chlorination of ilmenite, whole-rock fluorination, and electrolysis of molten lunar material, also have attractive aspects but are not as well studied as hydrogen or carbon monoxide reduction. Of these, whole-rock schemes have the distinct advantage that they require only minimal physical processing and do not require beneficiation and sizing of selected minerals. They also share a disadvantage: any scheme that produces a variety of metals requires a much more complex chemical processing plant to turn the metals into useful products. Chlorination and fluorination processes also require very efficient recovery of the halogen reactants from a wide diversity of halide products.

Nonferrous metals are also abundant on the Moon. Titanium from the rutile by-product of ilmenite processing is an obvious possibility for extraction. Aluminum from anorthite and magnesium from pyroxenes and olivine should also be considered. Some authors have advocated using lunar calcium (also from anorthite) for making electrical wires and cables for outside use in an environment free of water vapor and oxygen.

Perhaps the easiest source of ferrous metals is simple magnetic extraction of native metal grains from the regolith. These grains have two origins: nearly pure metallic iron, produced by reduction of FeO by implanted solar-wind hydrogen during impact shock-heating, and asteroidal iron–nickel–cobalt alloys left behind by the explosion of asteroidal impactors. Like polar ice and helium-3, these materials owe their presence on the Moon to external sources of hydrogen and metals: none are native to the Moon.

The use of lunar polar ice may be very difficult. It is stable only in permanently shadowed areas, so it is not easy to use sunlight to evaporate or melt the ice. If there are permanently illuminated mountain peaks adjacent to the ice deposits, one could envision a manned lunar polar base occupying that choice location, feeding solar power to a mining operation on the crater floor. However, even given such a fortuitous geographical arrangement, schemes for installing and operating a base and mine under such extreme conditions are very daunting. The installation of base modules requires landing in incredibly rugged terrain atop a lunar peak. Mining operations would have to be conducted at a temperature below about 100 K, a temperature so low that most metals would be brittle, and in the presence of pervasive fine, highly abrasive lunar dust. Ice at such low temperatures is as strong as rock.

Polar ice is not the only possible source of volatiles on the Moon. Solar-wind-implanted volatiles, especially hydrogen and helium, are widespread, especially in the equatorial region. These gases are implanted as ions in the surfaces of grains exposed atop the lunar regolith. Old, mature regolith has accumulated far more gases than young, immature regolith, from 10 to as much as 50 ppm (parts per million) by weight. The highest solar-wind gas concentrations are found in fine-grained ilmenite: it would be desirable to invent a beneficiation scheme that isolates ilmenite and screens out the largest grains. Unfortunately, mature regolith is very rich in agglutinates, which weld grains together irrespective of their composition, size, or gas content. Beneficiation of ilmenite or screening the ilmenite is impossible without thorough crushing to liberate individual grains. Further, the grains of greatest interest are so small that electrostatic "cling" makes it very difficult to separate them from grains of other sizes or compositions. Many simplified studies of ilmenite beneficiation have been reported; the most successful are those that least fit lunar conditions. The best recipe for success in liberating ilmenite is to separate magnetically, in air, a terrestrial ilmenite-bearing simulant free of agglutinates. But terrestrial ilmenite is a highly magnetic solid solution of iron oxides in ilmenite, not at all like lunar ilmenite. Success under such unrealistic conditions does not translate into success with unliberated, dust-like, slightly magnetic ilmenite in a vacuum.

The obvious method of extracting implanted hydrogen and helium is to heat the lunar material. It is reasonable that a feedstock containing 50 ppm of hydrogen could be heated efficiently so that 40 ppm of hydrogen is recovered. Then 25,000 tonnes of regolith must be heated per tonne of hydrogen recovered.

G.L. Kulcinski and his collaborators at the University of Wisconsin have suggested the recovery of solar-wind-implanted helium-3 from the lunar regolith for return to Earth for use as a clean fusion fuel with terrestrial deuterium (6). Near-perfect recovery of helium-3 from mature regolith would require heating 100,000,000 tonnes of regolith per tonne of helium-3 extracted. Based on extremely generous assumptions regarding helium-3 content, efficiency of gas extraction and recovery, beneficiation and sizing of ilmenite particles, heat reclamation from the baked regolith, sealing the ovens gas-tight in the presence of ubiquitous dust, and the ability of great quantities of complex mechanical equipment to survive autonomous operation in a dusty hard vacuum with diurnal temperature extremes of -200 to $+200^{\circ}\text{C}$, the scheme can be made to turn a handsome profit. Although it is far too early to identify this scheme as practical and economically attractive, it is not too soon to contemplate realistic tests of components of this proposed system in lunar simulators and even on the Moon itself.

Logistical Considerations. Outbound missions from low Earth orbit (LEO) to the lunar surface have total propulsive velocity requirements of 3.0 km s^{-1} (from LEO nearly to escape velocity) plus 0.7 km s^{-1} to match orbital speeds with the Moon and 2.4 km s^{-1} to decelerate the spacecraft to a soft landing on the lunar surface, a total of 6.1 km s^{-1} .

Inbound missions from the lunar surface to a LEO Space Station require the same total velocity change; however, because the Space Station is so close to the top of Earth's atmosphere, the 3.0 km s^{-1} of braking required for the returning lunar vehicle to match speeds with the LEO Station can be provided by aerobraking, the controlled passage of the space vehicle, with an appropriate

heat shield, through Earth's upper atmosphere below the 100-km altitude level. The perigee of the resulting capture orbit, ranging from about 60–100 km, must then be lifted by an apogee burn to circularize the orbit at the same altitude as the LEO station. This maneuver typically requires 0.3 km s^{-1} velocity change a total propulsive change of 3.3 km s^{-1} .

The Resources of Mars

In addition to the general books on space resources suggested earlier, Martian resource exploitation is a major concern at the Case for Mars Conference series. The proceedings of these conferences are the basis of a series of books starting with *The Case for Mars* (7–10). A popular account of planning for future manned Mars missions, including Mars resource exploitation, is found in the book, *Islands in the Sky* (11).

Because of its substantial escape velocity, Mars is even less suited for export of products than the Moon. Nonetheless, production of propellants on the Martian surface is highly attractive. Manned missions would also benefit greatly from extracting life-support materials, such as water, oxygen, nitrogen, and nutrients from local resources. A demand for volatiles naturally motivates a search for methods of extracting these materials from the Martian atmosphere, rather than from crustal minerals.

Materials Available on Mars. Geochemically, Mars is intermediate in properties between Earth and the Moon. The presence of a water-bearing carbon dioxide atmosphere on Mars enables a wide range of weathering reactions and products never found on the anhydrous, anoxic Moon. Solar ultraviolet photolysis of both carbon dioxide and water vapor provides a continuous weak supply of oxygen, which accelerates the weathering process and makes a variety of highly oxidized surface minerals, notably ferric oxides and oxysalts such as carbonates and sulfates. Much of the Martian surface has experienced the direct action of liquid water in the distant past. Clay minerals and hydrated salts are certainly widespread, possibly ubiquitous, components of the surface. Nitrogen, a minor constituent of the atmosphere, may also contribute small quantities of nitrate minerals by photochemical production of nitrogen oxides.

Primary igneous rocks on Mars have been spectrally mapped from Earth and from orbit. The discovery that the rare shergottite, nakhlite, and chassignite (SNC) achondritic meteorites contain adsorbed gases that have the distinctive chemical and isotopic signature of the Martian atmosphere has led to the general acceptance that they are fragments of the Martian crust that have been hurled into independent orbits around the Sun by asteroid or comet impacts on Mars. Calculations by Ann Vickery and Jay Melosh of the University of Arizona have shown that rock fragments from shallow depths in the regolith may be accelerated to escape velocity by the blast wave from such impactors (12). The SNC meteorites, of which more than a dozen have now been identified, are igneous rocks that contain elevated concentrations of ferrous iron and traces of weathering products characteristic of the Martian environment.

Direct chemical analyses of the Martian surface dirt have been carried out by the Viking landers and the Mars Pathfinder mission. The dirt is characterized

by high sulfur and iron and low potassium and sodium contents. This combination of factors is best understood as the result of oxidative weathering and hydration of a relatively low-temperature (FeO- and sulfur-rich) primitive material, followed by leaching and partial removal of water-soluble salts. In addition, traces of weathering products have been found in several of the SNC meteorites. Along with smectite- and illite-like clays, these products include sodium and potassium chlorides; hydrated ferrous oxides; and carbonates, phosphates, and sulfates of calcium and magnesium, some of which carry water of hydration.

The Martian poles are covered with deep, probably permanent caps of dust-laden water ice, often covered by frosts of solid carbon dioxide (“dry ice”). Because of the orientation of the line of apsides of the Martian orbit relative to the direction of the polar axis, the seasons are not symmetrical between the Southern and Northern Hemispheres: the colder (southern) polar cap presently retains solid carbon dioxide all summer. Seasonal precipitation of solid carbon dioxide (dry ice) snow occurs at high latitudes and altitudes. Ultraviolet spectra of the poles reveal the presence of a tiny trace of ozone in the ice deposit. The ozone is a minor product of atmospheric photochemistry.

Manufacture of Propellants on Mars. The first use of Martian resources is likely to be to produce propellants to support unmanned or manned missions to the Martian surface, including use for both local mobility and for return to Earth (Table 3). In the longer term, manned missions may use native Martian materials to provide their requirements of air (oxygen and nitrogen) and water.

By far the most accessible and most easily “mined” resource on Mars is atmospheric carbon dioxide. Ash et al., (13) first proposed that Martian carbon dioxide could be dissociated into carbon monoxide and oxygen by exposure to a hot (roughly 1000 K) zirconia membrane. By applying an electrical potential across the membrane, oxygen can be pumped selectively (as oxide ions), leaving a carbon dioxide/carbon monoxide mixture on one side and oxygen on the other

Table 3. **Resource Targets on Mars**

Product	Resource target	Process
Shielding	Regolith	Mechanical
Oxygen	Atmosphere	CO ₂ cracking—ceramic membrane electrolysis
		CO ₂ cracking—molten carbonate electrolysis
	Fe oxides; FeOOH	CO reduction
		H ₂ reduction
	Water	Electrolysis
Fe metal	Regolith	Any reduction process,
		Magma electrolysis
Water	Atmosphere	Condensation
	Clays	Heating
	Permafrost/polar ice	Melting, distillation
Refractories	Regolith	Electrolytic residue
Nitrogen/Ar	Atmosphere	CO ₂ removal

side of the membrane. The net effect is to dissociate CO_2 via



Liquefaction of CO and oxygen provides a propellant combination that has a specific impulse of about 284 seconds, limited by both the modest enthalpy of combustion of CO and the high molecular weight of the exhaust (carbon dioxide; MW = 44).

At high pressures and moderate temperatures, carbon monoxide can disproportionate via



to precipitate graphite, after which the carbon dioxide can be cycled through the zirconia cell to yield more oxygen.

A variety of additional propellants can be synthesized in the presence of other elements. If hydrogen is available, then hydrogen can be burned with oxygen ($I_{\text{sp}} = 460$), or hydrogen can be used with carbon oxides to synthesize any of a variety of other fuels, including methane (CH_4), methanol (CH_3OH), ethanol ($\text{CH}_3\text{CH}_2\text{OH}$), and acetylene (C_2H_2). These fuels burn with oxygen to provide a lower specific impulse than hydrogen–oxygen but have the great advantage that they are noncryogenic. Some are liquid at the normal temperatures of the Martian surface. A discussion of schemes for integrating hydrogen with carbon-oxygen chemistry is found in Hepp et al. (14).

Hydrogen gas is not found on Mars, so taking advantage of these options requires finding a source of hydrogen. The two most obvious alternatives are to transport liquid hydrogen to Mars from Earth or to extract hydrogen from Martian water. The former requires carrying liquid hydrogen for several months, far beyond present experience for space missions. Martian sources of water include atmospheric water vapor, water of hydration in surface clays and salts, permafrost (at latitudes greater than about 45°), and polar ice (above roughly 80° latitude). The atmosphere of Mars, even when saturated with water, is so cold that the water content rarely exceeds a few parts per million. Although processing a clean gas stream appears attractively simple compared to mining, the energy cost of removing so small a trace of water appears prohibitive (15). The most concentrated source of water is polar ice, but accessing that ice requires operating during local summer when solar energy illuminates and warms the polar cap above the winter temperature of 140 K.

Permafrost is so widespread on Mars that it may prove almost as attractive as polar ice; however, permafrost may be buried under 0.1 to several meters of relatively dry regolith (16). Removal of the overburden may be easy, but mining permafrost requires extreme care because permafrost is effectively an abrasive-ice composite material. Finally, we consider extracting water from heated regolith that contains hydrated salts and hydrated and hydroxyl silicates, which is apparently possible anywhere on Mars. This approach may be the most direct and simplest to implement (17).

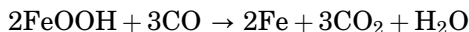
The second most attractive atmospheric component for propellant production is nitrogen. Nitrogen, combined with a chemical source for hydrogen and the ubiquitous carbon dioxide, allows synthesizing ammonia (NH_3), hydrazine

(N₂H₄), hydrazine derivatives, and the storable oxidizers nitrogen tetroxide (N₂O₄) and nitric acid (HNO₃).

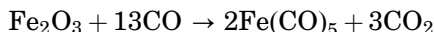
Life-Support Materials on Mars. Water and oxygen have obvious direct application to life-support systems on Mars. Inert gases, especially nitrogen and argon, are also of considerable importance because of their use in admixture with oxygen to moderate the danger of fire. Together, nitrogen and argon make up about 5% of the Martian atmosphere.

Nutrients for plants and animals derive ultimately from inorganic materials. Beyond those materials already mentioned, chemically active nitrogen compounds such as ammonium and nitrate salts and phosphates deserve special mention.

Metals on Mars. Because of the high concentration of iron oxides in both Martian weathering products and primary igneous rocks, extraction of iron seems attractive (Table 3). The route to iron reduction could be either by carbon monoxide or solid carbon; the former is simpler and more practical. Carbon monoxide heated strongly with an iron oxide-hydroxide weathering product releases both carbon dioxide and water vapor and produces pure metallic iron:



Recovery of water and cycling of carbon dioxide through a cracking device allows oxygen recovery and CO regeneration. Reaction of carbon monoxide with iron or iron oxides at about 200°C and pressures of the order of 100 atmospheres produces a volatile iron carbonyl:



These moderate temperatures can be attained by passive solar heating without requiring electrical heating. The gaseous iron pentacarbonyl can then be passed over a heated surface at about 200°C and 1 atm pressure to decompose it quantitatively into high-purity metallic iron and gaseous carbon monoxide. This is a form of *chemical vapor deposition*. Iron deposited in this way is so pure that it has the corrosion resistance of stainless steel.

No other metal should be as readily extractable as iron. The very existence of aluminum ores appears improbable; indeed, if the SNC meteorites are representative, even plagioclase feldspar may be of low abundance locally. Titanium ores are also not established. The iron titanate mineral ilmenite, found in high concentrations in several lunar mare basins, is a minor or trace mineral in SNC meteorites. Chromite is also found in trace amounts. Other sulfides occur in minor or trace amounts, too low for reasonable economic exploitation, but suggesting the possible presence of undiscovered sulfide ore bodies on Mars (18).

It is important to note that existing data on Mars are of insufficient resolution and sensitivity to confirm or reject the presence of ore outcroppings of other economically useful materials. Hydrothermal deposits, either active or extinct, are almost certainly present. Therefore, ore mineralization must be considered a very real possibility. However, without knowledge of the nature and location of these ores, it is scarcely possible to suggest methods of use. It must suffice to point out that volatiles and iron, as discussed earlier, meet our most

pressing local needs. Ancient sedimentary deposits might also provide calcium and magnesium carbonates, which suggest the possibility of manufacturing “Marscrete”, or Martian concrete, using carbonates that have been kilned to drive off carbon dioxide and make lime (CaO).

Logistical Considerations for Mars Missions. The total velocity requirement for outbound trips (from the LEO Space Station to the surface of Mars) is only 4.8 km s^{-1} , compared to 6.1 km s^{-1} for missions from LEO to the lunar surface. The reason for the advantage enjoyed by Mars is that aerobraking for arrival at Mars is possible, and permits the dissipation of hyperbolic orbital energy without significant propellant consumption. The penalty paid for this liberation from heavy propellant use is that a heavy aerobrake heat shield must be carried.

The escape velocity of Mars, 5.4 km s^{-1} , is the mean of the escape velocities of Earth (11.2 km s^{-1}) and the Moon (2.4 km s^{-1}). Single-stage-to-orbit launch vehicles, which are marginally feasible on Earth, are highly appropriate for takeoff from Mars. Because the most likely destination for a vehicle that departs from Mars is Earth and the entire impulse for injection into an Earth-intercept solar orbit can be delivered at launch while still deep in Mars' gravity well, summing velocity requirements is incorrect: energy conservation gives a return velocity requirement of 7.8 km s^{-1} for Earth intercept, compared to 3.0 km s^{-1} for aerobraked return to Earth or LEO from the Moon. The returning Mars vehicle must then use aerobraking either to descend to Earth's surface or to capture into Earth orbit.

Because of the relatively high escape velocity, it is probable that, aside from scientific samples and curios, export of Martian materials would be unprofitable.

The Martian Moons, Phobos and Deimos

The Martian moons are dark, irregularly shaped bodies whose characteristic dimensions are of the order of 10 km and masses of the order of 10^{18} grams (one trillion tonnes). Their orbits bracket Mars-synchronous orbit, so that Phobos passes from west to east across the Martian sky, whereas Deimos moves in the opposite direction. The masses and densities of these small satellites, measured by the Viking Orbiter spacecraft in the late 1970s, show that they are little more dense than ice and about half the density of even the least dense rocky meteoritic materials. Spectroscopic studies of the Martian moons have long suggested that they are very similar to the class of meteorites known as carbonaceous chondrites. Carbonaceous chondrites are black because of the pervasive presence of both tarry organic polymers and the iron oxide mineral, magnetite. They contain up to 6% of organic matter and up to 20% chemically bound water in the form of clays and hydrated salts. Spectroscopic studies in 1994, however, found that, unlike carbonaceous meteorites, Phobos and Deimos show no sign of the presence of water-bearing minerals.

One of the oddities of the Martian satellite system is its dynamic environment: the two satellites are so deep in the Martian gravitational field that most of the debris excavated from them by impacts remains in orbit about Mars. The debris fan from each impact quickly settles into a thin, dense dust disk in the equatorial plane, where it is eventually reaccreted by the two moons at low

relative speeds. In the Asteroid Belt, for comparison, debris is injected into independent orbit around the Sun. It is dispersed over a vast volume of space and is subjected to severe disturbing forces from radiative pressure and gravitational perturbations, so that there is no reasonable probability that it will be reaccreted by the body from which it originated—or any other asteroid. Therefore, the surfaces of Phobos and Deimos, more than any asteroid, consist of materials ejected and strongly shock-heated in many generations of previous impacts. The surface material, not surprisingly, is very dry. But the surfaces of these bodies do not tell us whether their interiors are rich in water. The low observed density likewise is ambiguous evidence because it could be affected either by a large abundance of water ice or by pervasive fracturing of the interior: the insides of Phobos and Deimos may be full of both ice and vast systems of cracks and voids.

Utilization schemes for Phobos and Deimos clearly must await determination of their water content. If water is present and abundant, both satellites would become highly attractive as sources of propellants for use within the Mars system and for return to Earth.

Logistically, because of their low surface gravities, Phobos and Deimos are both more accessible for outbound soft-landing missions than the Moon. Return missions from Phobos and Deimos to Earth intercept require a very small velocity change to escape from the satellite, a larger ΔV to escape from Mars into heliocentric orbit, and another substantial ΔV to lower the perihelion distance of that orbit to 1 AU. The total ΔV for return to Earth intercept (aerocapture) is 2.88 km s^{-1} from Phobos and 2.55 km s^{-1} from Deimos. To get off Phobos and drop down into the atmosphere of Mars for an aerobraked landing requires only 0.56 km s^{-1} and a Mars landing from Deimos needs only 0.4 km s^{-1} . Thus return to a Space Station orbit about Earth from either of the Martian moons is easier than a comparable return from Earth's Moon (assuming aerobraking in each case). The principal penalty for Phobos and Deimos return missions is the approximately 9-month travel time back to Earth.

The Near-Earth Asteroids

Asteroids as Threats and Opportunities. Recent developments in the astronomical search for near-Earth asteroids and in geological studies of impact features on Earth and other planets have presented us with a disturbing vision of the threat of disaster visited upon Earth by the impact of a near-Earth asteroid or comet. A technical overview of these issues can be found in the book *Hazards due to Comets and Asteroids* (19) and in a popular discussion based on that volume (20). A quantitative treatment of impact hazards, based on the results of extensive and detailed Monte Carlo simulations of centuries of impact events on a populated Earth, has recently appeared (21). These studies show that, if a large asteroid, one massive enough to threaten human civilization, is now on a course to collide with Earth 100 years from now, then that body would almost certainly be one that we have not yet discovered: to date, we have found only about 10% of all kilometer-sized bodies that cross Earth's orbit. These data also show that virtually all such large bodies can easily be discovered within the next few decades by a systematic, globally coordinated search and characterization program that costs

less than a single small space mission. Using such a search program to give us adequate warning of a threatened asteroid impact, we would then have ample time to design, build, test, and deploy an effective defense against the threat.

That same search and characterization program, however, provides an exceptional opportunity. Many of the most dangerous asteroids have orbits that are remarkably accessible from Earth: they cross or graze Earth's orbit about the Sun and can be reached, orbited, and landed upon more readily than any other body in the solar system. Fully a quarter of the near-Earth asteroids of all classes are easier to land on than the Moon: a given booster rocket could soft-land a larger payload on any of these than it could on the Moon. Further, these bodies have such feeble gravity that departure from them to return to Earth is vastly easier than departure from the Moon. Return of samples from the near-Earth asteroid Nereus, for example, requires a departure speed as low as 60 meters per second (135 mph), whereas departure from the Moon to return to Earth requires a speed of about 3000 meters per second. The amount of energy (or propellant) required per ton of returned material is 2500 times as large for lunar missions as it is for Nereus missions. For this reasons, exportation of bulk materials from the Moon to Earth could make sense only for fabulously valuable materials.

Logistical Considerations. Because of the ease of returning asteroid-derived materials to Earth orbit, very large masses of materials may be moved. Logistical studies suggest that each ton of equipment launched from Earth to a near-Earth asteroid can return 100 tons of material to Earth orbit during the operational lifetime of the vehicle. Thus, assuming a launch cost of less than \$500 per pound from the surface of Earth and using a 100:1 leverage factor, materials such as propellants or metals could be made available in Earth orbit for a few dollars a pound. The materials needed for future space transportation and construction would then be comparable in expense to those used in high-quality residential construction here on Earth.

In addition to these very favorable energy and logistical considerations, asteroids are attractive targets for a wholly different reason: they are rich in "cheap" materials, such as water or steel, that are of great value and utility in space but outrageously expensive to launch from Earth. Further, the large majority of near-Earth asteroids contains high concentrations of extremely valuable precious and strategic metals, such as platinum, osmium, iridium, rhenium, and palladium, and semiconductor components such as germanium, gallium, arsenic, antimony, tellurium, and indium. The Earth-surface market value of this fraction is roughly \$10,000 per pound, and the concentration of platinum, for example, is higher in the average NEA than in the best known terrestrial ore deposits. Experiments now underway in the laboratory of Prof. Henry Freiser at the University of Arizona, funded by a visionary private foundation that is interested in commercial development of space, are seeking simple, effective means to extract and separate many of these valuable resources from meteorites that are authentic samples of asteroids.

The idea of extracting asteroidal materials for commercial use is not new: As early as 1903, the great prerevolutionary Russian rocket visionary, Konstantin Tsiolkovskii, proposed exploiting asteroid resources. The father of practical rocketry, the American physicist Robert Goddard, wrote in a 1918 essay entitled *The Ultimate Migration*, of interstellar ships made from asteroids conveying our

remote descendants away from the death throes of the Sun. The idea of mining asteroids was so visionary that Goddard sealed his manuscript away in an envelope labeled "Special formulas for silvering mirrors," where it languished unread for more than 60 years. Goddard's reticence is understandable: the technologies required to carry out such ambitious schemes did not exist then, whereas critics and mockers incapable of understanding his ideas were legion. But today, this dream can be made a practical reality by applying the technology of the year 2000.

The keys to successful importation of materials from space are lower launch costs, careful choice of exploitation targets to favor those that are most accessible and have the richest resource concentrations, and minimizing the complexity of the operations to be undertaken by mining and extraction vehicles to rely on artificial intelligence, not human presence. The question how to lower the cost of access to space from about \$5000 per pound to a few hundred dollars per pound is also an old one. Exactly a century ago, Tsiolkovskii wrote a science fiction novel in which the first successful manned venture into space was carried out in the year 2000... by a consortium of industrialists, scientists, and technologists funded by what can only be described as venture capital. Today, we see nearly two dozen companies, funded by venture capital and equipped with exciting ideas and the most modern aerospace and electronic technology, competing to lower the barriers in the way of massive development of space.

The NASA Near-Earth Asteroid Rendezvous mission, which arrived at the asteroid Eros in 2000 for a prolonged study of its surface, has already returned a wealth of data on the physical and chemical properties of that asteroid. Early data strongly suggest that the common S-type asteroids are chemically closely similar to ordinary chondritic meteorites, the most common class falling on Earth. X-ray and gamma-ray analytical experiments are planned to continue compositional mapping of Eros for many months.

The Space Development Corporation's much-discussed Near-Earth Asteroid Prospector mission (NEAP) is merely the first example of attempts to privatize space activities. For planning purposes, it is assumed that NEAP is a mission to the asteroid Nereus. NEAP, although confined to a role as a science platform to fly NASA-funded instruments at low cost and to sell data gathered by privately funded instruments to NASA, can be regarded as an authentic prospector—a searcher for useful resources. The leading candidate instruments for the first mission are an alpha-proton X-ray spectrometer and multispectral imaging for global physicochemical mapping of the surface. The NEAP spacecraft bus can also accommodate several small experiments, such as a rover, that could be soft-landed on the surface of the asteroid after the primary mapping mission is complete. A second step would be to land processing experiments on the surface of an asteroid and demonstrate small-scale production of water or metals. Because return of materials from so many NEAs is exceptionally easy, both virgin surface samples and processed materials may be automatically returned to Earth more easily than samples could be returned from the Moon (a technique the Soviet Union that demonstrated successfully on the unmanned Luna 16, 20, and 24 missions in the 1970s).

Given the accessibility of NEAs and the diversity of their resources, some idea of the amount of available resources is in order. Perhaps the first example of

note is the metallic near-Earth asteroid called Amun. This asteroid is about 2000 meters in diameter. If it were to strike Earth, it would deliver a devastating blow of 10 million megatons (10 teratons) of TNT, several thousand times the explosive power of a nuclear world war. Amun, the smallest known metallic asteroid of the several dozen known, contains several times as much metal as the entire amount of metals mined and processed during the history of humankind. A conservative estimate of the market value of this asteroid is \$5 trillion.

Resource Richness and Diversity. The entire NEA population, which has very diverse chemical and physical properties, contains vastly more material than Amun. An estimate of the overall composition of the NEA population is shown in Table 4. It is not difficult to estimate how much of each of a wide variety of commodities, such as water, carbon, nitrogen, metals, and phosphorus, is required in circulation to maintain one average human at present-day North American, Western European, or Japanese levels of affluence. From these figures, we may estimate how many people could be supported indefinitely by the resource wealth of the NEA population, assuming a fully recycling regime powered by the Sun. According to Table 4, the number is probably close to 14 billion people. Nitrogen, principally in its role as a fire-suppressing diluent of atmospheric oxygen, appears to be the limiting resource. One of the most remarkable lessons of Table 4 is that the proportions of materials needed by civilized human beings is similar to the proportions in asteroids.

But the NEAs are only a small part of the picture. Most NEAs follow orbits that take them out to the heart of the Asteroid Belt, between the orbits of Mars and Jupiter, at aphelion. Thus a processing unit landed on a typical NEA will get

Table 4. Resources of the Near-Earth Asteroids^a

Commodity	Mass among NEAs, 10 ¹⁵ g	Per capita inventory, g/person	Population sustainable by NEA resources, billion people
Silicate	2500	140,000,000	17.8
Ferrous metals	300	20,000,000	30.0
Fe in oxides	300		
Cement	60	10,000,000	6.0 ^b
Industrial CaO		2,000,000	30.0
Phosphates	10	2,000,000	5.0 ^c
Water	300	10,000,000	30.0
Carbon	100	1,000,000	100.0 ^d
Nitrogen	10	700,000	14.0 ^e
Sulfur	60	1,200,000	50.0
Sulfides	150	1,200,000	125.0

^aThe near-Earth asteroid population is a renewable resource that replaces itself about every 30 million years. Note that the only true “consumable” in this fully recycling system is solar power.

^bCement is of dubious utility in space, except for wholly internal (non-vacuum-tight) construction.

^cPhosphate fertilizer usage on Earth is predicated upon toleration of massive loss in runoff from fields. In a 100% recycling regime, the required inventory could easily be 10 times smaller.

^dCarbon inventories assume 1000 g of plant carbon per gram of human carbon.

^eNitrogen inventories assume 1000 m³ of habitat volume per person.

a free round trip to the Belt and back on each trip around the Sun (typically, once every 3–5 years).

In the likely order of development, water is the first asteroidal resource worthy of attention. Its uses as a propellant and as a life-support material are obvious. Second in order would be native ferrous metal alloys, whose major components, iron and nickel (and possibly cobalt), would be retained in space for constructing space-based facilities such as solar power satellites. The rare and very valuable precious metals and semiconductors in asteroidal metal alloys are worth returning to Earth. Dr. Jeffrey S. Kargel has explored the effects of large-scale importation of these materials from space on the market size and prices of these commodities on Earth. He concludes that prices will decline less rapidly than the rate of supply increase because of new uses stimulated by lower prices.

The potential customers for the materials that remain in space include government agencies that need propellants for injection into geosynchronous orbit, for orbital stationkeeping, or for departure from Earth's gravity well, such as both unmanned and manned Mars missions. In addition, civil traffic bound for geosynchronous orbit would be an important market for propellants. Metals would be of use for constructing and shielding large structures, of which the most obvious commercial example would be Solar Power Satellites. But this entire scenario depends on a stable and rational legal and regulatory system in which investors will have reasonable assurance that the fruits of their ingenuity and investment will not be arbitrarily obstructed, or even confiscated. We return to this issue later.

Proposed System Architectures. There are two central questions regarding the exploitation of asteroidal resources. First, there are the propulsive energy constraints imposed by the geometry of the orbit of the target asteroid. Outbound ΔV requirements from reference orbits such as LEO (a circular orbit at 300–600 km altitude, compatible with future space stations) have been calculated for almost every known near-Earth asteroid. In general, the typical NEA of interest has a perihelion close to or within Earth's orbit. Many NEAs have aphelia in the asteroid belt. The best candidates have low orbital inclination, perihelia (or, in the case of Atens, aphelia) very close to 1.00 AU, and low eccentricities. Outbound ΔV s (from LEO to landing on the target asteroid) as low as 3.2 km/s are allowed by theory, but in practice bodies whose outbound ΔV s are less than about 4.2 km/s are so vulnerable to perturbations and capture by Earth that their populations are depleted. About 15% of all NEAs have outbound ΔV s that are less than the ΔV requirement for missions from LEO to a soft landing on the Moon (6 km/s). Return from many NEOs requires ΔV s less than 1 km/s; several lie below 0.4 km/s. The best known return ΔV is 0.06 km/s from Nereus (1982 DB). By comparison, return to Earth intercept from the lunar surface requires roughly 4 km/s, depending on where the launch site is on the Moon.

Second, there is the choice of location for the space base in near-Earth space from which missions depart and to which they return with their cargoes. In a study of opportunities for missions to the Mars system, Benton Clark argued that the most suitable base for long-term, round-trip use would be in highly eccentric Earth orbit. A typical base orbit might have a perigee of several thousand kilometers, beyond the heart of the Van Allen radiation belts, and an apogee of

40,000 to as much as 400,000 km (roughly the orbit of the Moon). Both aerobraking and propulsive capture require that the hyperbolic excess velocity be lost as close to Earth as possible and the perigee be about 100 km. A small apogee burn can then lift the perigee safely out of the atmosphere. From such an orbit, the velocity required for escape from Earth is extremely modest, and from such an orbit, it is easy, by a brief apogee burn, to drop the perigee down to the top of the atmosphere either for return to Earth or for an engine burn to escape Earth at maximum efficiency.

Solar Power from Space

In recent years, debate concerning the greenhouse effect, global warming, fossil fuel burning, and the environmental impact of mining and transporting fossil fuels has focused increasing attention on alternative sources of electrical power. The desire is for cheap, clean, abundant future energy supplies motivates consideration of solar power.

There are several ways in which solar power may be made available on Earth. The first and most obvious option is constructing solar voltaic cell “farms” in locations on Earth’s surface where the total insolation is highest. Earth-surface solar cell arrays are subject to the day–night cycle; atmospheric attenuation of sunlight by Rayleigh scattering and by cloud opacity; shadowing by windblown dust (an especially common phenomenon in areas that have the most reliable solar illumination); or by smoke, haze, or volcanic aerosols, and damage by wind, snow and ice loads, hail, lightning, and flying objects.

An obvious alternative is to locate the solar cell arrays in space, where performance (watts of electrical power per square meter of collector area per year) can be 2 to 10 times higher than on Earth’s surface (22). The principal drawbacks of space-based solar collectors are the initial cost of launch and the necessity of beaming the power down to receiving antennae on Earth as microwave power. The choice of the orbit of the solar cell array has an important influence on system performance. The easiest orbits to achieve (low Earth orbit; LEO) usually have a 50% duty cycle due to passage through Earth’s shadow and also traverse the 300 to 1000-kilometer altitude range where the probability of collision with orbital debris is high. Higher orbits have greater installation costs and less opportunity for service but better exposure to the Sun and a more favorable debris environment. Most point designs of Solar Power satellites (SPS) assume that they will be installed in geosynchronous orbit. Large-scale reliance on SPS constellations would entail installing approximately $5 \times 10^4 \text{ km}^2$ of solar collectors in the geosynchronous belt, where they would far outshine the brightest stars.

An interesting variant of the SPS concept that would avoid the light pollution problem is David R. Criswell’s scheme of building solar-power collectors on the Moon, using lunar resources for collector construction (23). This approach saves the cost of lifting most of the mass of the SPS system from the Earth.

SPS constellations in Earth orbit could also benefit greatly from using nonterrestrial materials. A recent NASA study (24) showed that SPS are close to being economically viable, even with Earth launch (“uphill” transport) of all of

their components. Any scheme that derives the most massive, low-tech SPS components from nonterrestrial metals (brought “downhill” to Earth orbit) holds promise of making them highly competitive economically.

Some words are also in order regarding the environmental impact of space resource use. In general, any industrial activity that can be off-loaded from Earth eases the environmental burden on Earth’s biosphere. An enormous proportion of Earth’s environmental troubles are related to mining, refining, transporting, and using fossil fuels. National policies aimed at reliance on solar power satellites mitigate all of these problems, simultaneously making the United States the world’s largest exporter of energy and energy technology and extending the usefulness of our limited petroleum reserves far into the future by dedicating crude oil to petrochemical production, not combustion. Finally, such an approach reduces or eliminates American dependence on foreign sources of crude oil.

Long-Term Prospects

For logistical and programmatic reasons, the most attractive locations for resource extraction are the near-Earth asteroids, the Moon, and the Mars system, including Phobos and Deimos. But, in the longer term, other options may be considered. As mentioned earlier, Mercury’s polar ice deposits provide a potential source of propellants for a sample return mission. Venus, by reason of its enormous atmospheric pressure (92 atmospheres), scorching surface temperatures (750 K, or about 900°F), highly corrosive atmosphere (containing gaseous hydrochloric and hydrofluoric acids and clouds of sulfuric acid droplets), and high escape velocity (10 kilometers per second), is among the most unattractive locations in the solar system, similar to Dante’s vision of Hell.

Near-Earth asteroids typically reach the heart of the Asteroid Belt at aphelion. Any equipment riding on an NEA is transported out to the main Belt every few years. It is easy to envision a device riding on a water-bearing NEO out to the Belt, manufacturing propellants from the water as it goes, and using that water to transfer from the NEO to a convenient Belt asteroid. Such a transfer is energetically no more difficult than returning to Earth. Emplacing equipment on Belt asteroids in this manner is logistically more complex than landing on an NEO but not more demanding from a propulsive or launch-weight perspective. In effect, systematic exploitation of the NEO population makes access to the main Belt easy.

The mineral riches of the Asteroid Belt are almost beyond comprehension (Table 5). The asteroids that reside in the Belt make up a very large population, roughly 40,000 bodies larger than a kilometer in diameter, and the total mass is roughly a million times as large as the total mass of the NEA population at any one time. Table 5 summarizes the total amount of resources available in the Belt in a manner similar to the NEAs treated in Table 4. The conclusions are staggering: the materials available in the Asteroid Belt would maintain indefinitely a human population of at least 10,000,000 billion people—about one million times the maximum carrying capacity of Earth. Assertions that we are running out of resources reckon without twentieth and twenty-first century technology. The supply of resources available to a spacefaring humanity is effectively infinite. But

Table 5. Resources of the Asteroids Belt^a

Commodity	Mass in the belt, 10 ²¹ g	Per capita inventory, g/person	Population sustainable by belt resources, billion people
Silicates	2500	140,000,000	17,800,000
Ferrous metals	300	20,000,000	30,000,000
Fe in oxides	300		
Cement	60	10,000,000	6,000,000 ^b
Industrial CaO		2,000,000	30,000,000
Phosphates	10	2,000,000	5,000,000 ^c
Water	300	10,000,000	30,000,000
Carbon	100	1,000,000	100,000,000 ^d
Nitrogen	10	700,000	14,000,000 ^e
Sulfur	60	1,200,000	50,000,000
Sulfides	150	1,200,000	125,000,000

^aAssumes full recycling and full reliance on solar power.
^bCement is of dubious utility in space, except for wholly internal (non-vacuum-tight) construction.
^cPhosphate fertilizer usage on Earth is predicated upon toleration of massive loss in runoff from fields. In a 100% recycling regime, the required inventory could easily be 10 times smaller.
^dCarbon inventories assume 1000 g of plant carbon per gram of human carbon.
^eNitrogen inventories assume 1000 m³ of habitat volume per person.

there is no possibility of returning all that material to Earth (enough steel in the Asteroid Belt, for example, to build a steel-frame building 8,000 stories tall covering all the land area of Earth), nor is there any possibility of accommodating so many people on one planet. The bulk of these resources would be used in space.

Beyond the Belt, in two extensive clusters located on the orbit of Jupiter, 60° ahead of and behind Jupiter, are the Trojan Asteroids. The total mass of material in these two vast clouds is probably several times the total mass of the Belt. At the same distance from Earth are the satellites of Jupiter. The largest four, the Galilean satellites Io, Europa, Ganymede, and Callisto, lie so deep in Jupiter's gravity well and so far inside its intense radiation belts that they are demanding destinations for even flyby spacecraft. The four small inner satellites are in an environment that is essentially not survivable. The two families of small outer satellites, however, are dynamically and probably even chemically indistinguishable from the Trojan asteroids. Saturn also has several small, remote satellites. But the great distance of these bodies from any plausible site of demand for their resources and the very low light levels at such great distances from the Sun (27 times less at Jupiter than at Earth) argue against any short-term economic significance.

The gas-giant outer planets, Jupiter, Saturn, Uranus, and Neptune, offer astronomical quantities of attractive resources situated at the bottom of very deep gravity wells. The easiest of the giant planets to land on and return from, Uranus, is at the very limit of the capability of single-stage nuclear rockets. But if high-performance liquid- or gas-core nuclear rocket engines, or fusion rockets, can be built, then the helium-3 fusion fuel resources of their atmospheres would become accessible to Earth. Uranus alone contains enough helium-3 to support a

population of 10 billion people at present-day North American or European consumption levels for more than 10^{15} years, or to support a million times as many people from now until the Sun dies of exhausting its hydrogen fuel.

Legal and Treaty Issues Governing Space Resource use

It is not sufficient merely to summarize the scientific data on the nature and orbital properties of potential space resources and outline proposed extraction, processing, fabrication, and transportation schemes attendant on their use. It is also necessary to survey the legal issues involving claims of ownership, national sovereignty, and mineral claim registration and certification.

Present Legal Regime. The booklet *Agreement Governing the Activities of States on the Moon and Other Celestial Bodies* contains the texts and history of a number of international treaties and agreements regarding space resources (25). The original agreement, and hence the basis for all more recent treaties, was laid by the 1967 *Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies*, usually referred to as the Outer Space Treaty. Article I states that

The exploration and use of outer space, including the moon and other celestial bodies, shall be carried out in the interests of all countries, irrespective of their degree of economic or scientific development, and shall be the province of all mankind.

It further states that the exploration and use of celestial bodies shall be done “without discrimination of any kind.” Article II proclaims, “Outer space, including the moon and other celestial bodies, is not subject to national appropriation by claim of sovereignty, by means of use or occupation, or by any other means.” Article VI binds each signatory State to take responsibility for “national activities in outer space, including the moon and other celestial bodies, whether such activities are carried on by governmental agencies or by non-governmental entities...” Article VII assigns liability for damage by space missions to the originating State, a statement of the principle embodied in the 1972 *Convention on International Liability for Damage Caused by Space Objects*. Article XII requires that all “stations, installations, equipment and space vehicles on the moon and other celestial bodies shall be open to representatives of the other States Parties to the Treaty” and to other States on the basis of reciprocity. Article XIII places intergovernmental agencies under the same Treaty obligations as individual States.

In general, the Outer Space Treaty assumes permanent domination of space activities by governments, a phenomenon that accurately characterized the 1960s, and further assumes that commercial ventures could be dismissed without individual consideration by simply assigning them to the nearest available government. Private property is not even mentioned. The 1967 Treaty was ratified by every spacefaring nation and reflects a very broad international consensus. For our purposes, the most important single feature of the Outer Space Treaty is the assertion in Article I that uses of space are “the province of all mankind.” The interpretation of this phrase is almost infinitely varied, reflecting

the political view of the interpreter. The most basic interpretation is that everyone has the right to participate in exploration and use of space and no one can be denied that right; the most ambitious would require representation of (or permission from) every nation on every mission. Acceptance of the Treaty by a wide range of nations that have different and often contradictory political and economic ideologies suggests that agreement on its wording simply masks disagreement on its meaning and application.

The most relevant document concerning exploitation of asteroids is the 1979 *Agreement Governing the Activities of States on the Moon and Other Celestial Bodies*, usually known as the Moon Treaty. Draft materials for this treaty and an account of the negotiations that led to it are included in a U.S. Government Printing Office publication, 59-896 O (25). The final agreement, adopted and opened for ratification by States by General Assembly Resolution A/RES/34/68 bearing the same name, consists of 21 articles.

Article 1 states that all other celestial bodies and the Moon are covered by this agreement, excepting meteorites that fall to Earth “by natural means.” Article 7 prohibits “the disruption of the existing balance of the environment” and calls on States to avoid “harmful contamination” of the body as well as “harmfully affecting the environment of Earth.” Article 8 guarantees the right to land, establish bases, and travel over the surfaces of these bodies. Article 9 enjoins these States to “use only that area that is required for the needs of the station.”

Article 11, by far the most important for our purposes, states “the Moon and its natural resources are the common heritage of mankind.” Further, “the moon is not subject to national appropriation by any claim of sovereignty, by means of use or occupation, or by any other means.” Neither the surface, subsurface, nor natural resources of the Moon “shall become property of any State, international governmental or non-governmental organization, national organization or non-governmental entity or of any natural person.” Installation of equipment or bases does not create any claim of ownership.

The States Parties to this agreement hereby undertake to establish an international regime, including appropriate procedures, to govern the exploitation of the natural resources of the moon as such exploitation is about to become feasible.

States Parties are required to report any natural resources discovered on any celestial body to the Secretary-General of the UN. Of course, in reality, the astronomers who examine the spectra of asteroids and find that they are made of metals or hydrated minerals are completely unaware of this treaty obligation and of the possible role of these materials as resources. Therefore, none of the roughly 1000 spectrally characterized asteroids has ever been “reported” to the Secretary-General. In closing, Article 11 proclaims that

The main purpose of the international regime to be established shall include: a) the orderly and safe development of the natural resources of the moon, b) the rational management of those resources, c) the expansion of opportunities in the use of those resources, and d) an equitable sharing by all States Parties in the benefits derived from those resources, whereby the interests and needs of the developing countries, as well as the efforts of those countries which have contributed either directly or indirectly to the exploration of the moon, shall be given special consideration.

Article 12 allows ownership of devices and stations placed on the Moon, and Article 14 places full responsibility for all activities on the Moon upon the State from which that activity originates. Here again, the possibility that commercial entities will operate in space is simply ignored. Article 15 requires that every installation on the Moon be open to inspection visits from any other State, and the remaining Articles refer to the role of international governmental organizations and to the ratification and amendment of the Agreement.

It is clear that the nations that contributed to this document accept absolute centralized control, common ownership of these resources, and required “sharing” (confiscation) of much of the proceeds of the work. The treaty seems to offer nothing for any State that can conduct operations on the Moon.

The United States and essentially every other nation that has spacefaring capabilities are signatories to all of the other treaties and agreements cited; however, with almost equal unanimity, these same nations have failed to ratify the Moon Treaty. Many other nations have ratified it, hoping perhaps for windfall profits from the labor, investment, and invention of others.

Suggested Legislative Actions to Support Space Development. Perhaps it would be more fruitful to ask what legal, regulatory, and economic regime would permit development of space resources. Which matters should be subject to international regulation, which functions should be carried by national governments, and which should be left to the free will and choice of the private and public spacefaring parties themselves? It appears that the appropriate role of government should be defined. We must acknowledge that opening the space frontier, like opening the American West, can be assisted or hindered by governmental actions. Just as in the establishment of railroads in the West, government assistance, not domination, is useful in several areas. The first four areas are of general applicability to space development:

1. Governments can support private space endeavors by buying scientific data, in effect, *privatizing many research missions*. The Near-Earth Asteroid Prospector (NEAP) mission is but the first example of this approach and offers a cost structure that reflects competitive commercial practice rather than the inefficiency that is characteristic of monopolies. NEAP and its likely near-future competitors will offer prices about a factor of 4 lower than the cost of doing a comparable mission from within the government. This improvement is above and beyond the already substantial cost reductions brought about by NASA under the administration of Mr. Goldin. This goal will largely be achieved in the United States under the provisions of the Commercial Space Act of 1997.
2. Governments should support the development of low-cost space transportation systems by *buying launch services competitively from private vendors*. A good start on this initiative was made in the United States by the Launch Services Purchase Act of 1990.
3. Governments, through NASA, ESA, Glavkosmos, or equivalent agencies, should take a leading role in *developing key electronics technologies* to assist early economic development of space, as in the early days of experimentation with communication, navigation, earth resources, and weather satellites.

4. Governments should play a major role in *developing critical propulsion technologies*. At a bare minimum, solar thermal and solar sail propulsion systems should be developed and placed in the public domain. The Russian test of the *Znamya* steerable solar mirror in space in November 1998 was relevant to this capability.

Three more goals are strongly specific to space resource development:

5. Governments should take the lead in *developing technologies for extracting and processing the most important mineral resources*. The technologies that should be tested in microgravity at the International Space Station, or elsewhere, include crushing and magnetic separation for extracting of native ferrous metal alloys and extracting water from ice lenses, permafrost, hydrated salts, and clay minerals. Initially, water may be used directly in solar thermal or nuclear thermal rockets, but it is clear that, in the longer term, converting water into cryogenic propellants will be highly desirable. Electrolysis of water into hydrogen and oxygen and liquefaction of both gases to make liquid rocket propellants must also be adapted to operation in very low (or artificial) gravity. Many of these technologies are extensions or adaptations of familiar Earth-surface processing technology to high-vacuum, microgravity environments. Metals mining and processing concerns have extensive Earth-related experience but require expert assistance from universities and government research centers in selecting their targets and from the government and the aerospace industry in adapting their experience to space operations. Given the present precarious financial state of the aerospace industry, its participation in the experimental stage of such an endeavor would almost certainly require government support. As recently as 1993, NASA was spending more than \$2.5 million per year on research into the use of nonterrestrial resources. This endeavor has been heartily endorsed by the NASA Administrator in many public addresses, but nonetheless, funding has essentially disappeared.
6. Governments should take the lead in *purchasing products produced in space for use in space*. The leading example is rocket propellants: the costs of ambitious deep-space missions are raised enormously by the cost of lifting the required fuel load for the outbound trip out of Earth's gravity well—at a cost of several thousand dollars per pint. Instead, all outbound missions that pass through low Earth orbit could be refueled with high-performance asteroid-derived liquid hydrogen and liquid oxygen from a “gas station” in low Earth orbit, perhaps an adjunct to the Space Station. Because the government would simply purchase high-performance, cheap, space-derived propellants that are competitive with Earth-launched propellants, this activity would also save the taxpayer money. Up-front investment in experimenting with the transfer of cryogenic propellants in microgravity would be required before any benefits could be realized. It is interesting that the Soviet Union has used in-orbit propellant transfer routinely and without incident since the 1970s, whereas the United States has never developed the capability. Russian experience would be very useful in this area. Beyond propellant manufacture, the most promising single space resource for immediate exploitation is native ferrous metals, which

could be used for large-scale space construction. Using asteroidal metals to build Solar Power satellites appears to be the easiest route to energy self-sufficiency for the United States, Japan, and Europe.

7. Finally, in the light of the economic potential of the points discussed before, governments must be prepared to deal with *certifying private claims of ownership and mineral claims on bodies in space*. Each national government could establish a formal means of registering these claims that will assure entrepreneurs that the integrity of their claims will be recognized, a necessary precondition to raising large amounts of venture capital. Governments interested in space development will predictably avoid international legal entanglements such as the original Law of the Sea treaty that have a confiscatory attitude toward profit and new technologies. Such treaties prevent the development of resources that would ease the lot of all humankind. Developing nations must understand that opening up new resources and reducing costs is to their direct benefit, that they too may directly participate in such activities. As pointed out earlier, the resources available in nearby space are so large and the cost of access to space will soon be so low, that the idea of domination by any monopoly is absurd. The most important effect governments can have on space development is to lay a firm foundation for commercial, competitive, private development of space resources.

Possible International Regimes. For the purposes of the following discussion, we will assume that some profitable form of space-resource enterprise exists or demonstrably could exist. If this were not true, none of the preceding questions would be of any interest. "Profitable" means that the proposed space activity can either provide a commodity in some economically meaningful location at lower cost than its competition, provide much larger amounts of that commodity at comparable prices, or provide a new commodity not already on the market. In all of these cases, it is to the economic advantage of the buyer, the entrepreneur, and the governmental entities that tax profits to allow such an activity to flourish. This is true even without allowing for the environmental advantages of relegating energy and mineral mining to space, which positively impact every person on Earth. It is perfectly true that certain existing interests, especially those that have monopolies on rare materials, would be harmed by such competition based on new resource bodies and new technologies. But this has always been true and always will be. Successful mining concerns are those that have taken the lead in adopting new technologies and developing new ore bodies. Metals concerns such as INCO will face the choice of trying to compete with space-derived resources or of taking a leadership role in their development. All consumers will benefit; only those corporations ready and willing to adapt will benefit from this new trillion-dollar market.

Therefore, we require that some means be instituted to make space mining possible, on the grounds that it is advantageous to humanity. This in turn requires that the entities that mine, extract, and fabricate space resources be given a regulatory regime in which investments would be rational. There are enough economic and physical risks associated with space mining that adding the risk of

an unstable or politicized regulatory environment would be fatal. Therefore, it is essential to register, recognize, and enforce mining claims.

The simplest, and in many ways most attractive, scheme would have the United Nations serve as a registry for mineral claims, much in the same way that it maintains a registry of launchings of space vehicles without exercising any control or authority over the activity. The World Court could also provide a venue for claim registry. Another method, inspired by patent law, would allow private individuals or corporations of any nation to register mining claims with their own government, with full mutual recognition of claims. In practice, nations could enter this arena one or two at a time, executing reciprocal agreements with the other relevant governments, as the need arises. Adjudication of conflicts might fall within the venue of the World Court or national courts. The World Court's role is limited by the historical absence of any enforcement ability. Enforcement would, of course, be carried out on Earth: the science-fiction device of asteroid miners declaring autonomy is not a near-term option, so long as space activities are highly dependent on Earth for both equipment and markets.

Several levels of presence might serve as the threshold for making a valid mining claim:

1. at the lowest level, discovery of an asteroid;
2. remote spectral characterization of an asteroid demonstrating the presence of an economically attractive resource (i.e., an ore);
3. physical presence of an unmanned vehicle to document the presence and setting of an ore;
4. physical presence of an unmanned vehicle, documented by sample return to Earth;
5. physical presence of a human crew that proclaims mineral rights or ownership; and
6. presence of an established human settlement.

It is noteworthy that, of the roughly 15000 asteroids discovered and cataloged to date and of the nearly 1000 that have been subjected to photometric or spectroscopic study, not one single astronomer has ever registered a public claim of mineral rights or of ownership. This is *prima facie* evidence for a universal consensus that such data do not constitute valid grounds for a claim. Maritime law has provided a precedent for claiming "mining" rights to shipwrecks through reconnaissance by unmanned vehicles that visit and document the "ore." Some space law theorists have suggested awarding a higher level of claim to those who return a sample to Earth. However, there seems to be no clear legal precedent for this requirement, and a host of scientific and engineering reasons to suppose that physicochemical characterization by spacecraft instrumentation provides all of the essential information that a returned sample could provide. The need to assay for the abundance of a particular ore, an economic necessity in mineral claim assessment on Earth, is largely irrelevant on homogeneous asteroids.

For this reason and others, there seems no need to require human presence because the evidence needed to establish the presence of an ore can readily be acquired without human presence. Further, requiring human presence places an

enormous economic barrier in the way of asteroid resource exploitation and effectively prevents private corporations and small nations from participating. Only a handful of superpowers could afford to carry out such a mission, artificially turning space resource exploitation into a monopoly or elitist activity. This is precisely the opposite of the desirable regime, in which many nations and companies can participate and in which free competition places constant pressure on prices, to the benefit of consumers. Note that manned presence on another celestial body, the Moon, has not resulted in any claim because the entity that carried out the Apollo project was NASA, a government agency that, by recognized international law, cannot make a claim of national sovereignty. The suggestion that the United States Congress should pass a law authorizing an "extraterrestrial land claim made by any private entity that has established a true space settlement" is completely inadequate. This is equivalent to a home builder not being allowed to seek ownership of the land on which the house was built until after he had built his house and moved in. In a commercial regime, investors need reasonable prior assurance that their capital is not being squandered. Claim recognition *must* precede development.

In general, the vast amount of resources available among the near-Earth asteroids and in the Asteroid Belt suggests that competition for mining claims should not be a common problem. Nonetheless, for any particular resource such as water, there may be a single asteroid that is significantly more accessible than others. There would then be a strong incentive to get there first and file a mineral claim that included the entire asteroid. For a 100-meter or 1000-meter body, comparable in size to an open-pit mine on Earth, it would be perfectly reasonable for the first arrival to claim the entire body. On the Moon, filing a mineral claim for, say, the entire Mare Imbrium, would be an absurdity, comparable to claiming all of Colorado or Switzerland as a mine site. Further, on a very small asteroid, minor environmental disturbances caused by one mine (dust, for example) might materially hinder other activities on the same body. Although it is true that the asteroidal environment is rapidly self-cleaning (through reaccretion of dust and Poynting–Robertson removal of escaped dust), the short-term local effects on visibility may be serious. A reasonable criterion might be to allow claims of all of the material of that body within 1 kilometer of the designated mine site in all directions, excepting any material already claimed in conformance with this criterion. Therefore, documentation of the claim must include physicochemical data on the surface and also a good three-dimensional map of the body where the mine site is specified. An actual landing at the mine site (not necessarily permanent occupancy) may be mandated, although there seems to be no convincing reason to require it. The concept of an ore body has quite a different meaning on an asteroid from usual terrestrial experience. Except for composite asteroids assembled by chance low-velocity collisions, most NEAs should be compositionally uniform. The physical state (dust, sand, cobble, boulder, country rock) of the resource is usually a more important determinant of mine-site location than the composition, which will usually be very uniform. There will rarely be veins of ore to follow. Further, the dross from extracting any resource (such as water from a carbonaceous asteroid) will itself be a very valuable resource, containing an abundance of other volatiles and both ferrous and precious metals.

This brings us to the issue of actual private or corporate ownership. A valid claim confers essentially all of the benefits of ownership except, perhaps, the ability to sell or license the claim. This author would regard a claim that can be sold as the moral equivalent of property. Allowing claims to be sold permits optimizing expertise in prospecting, site-study, and mine-development specialists. Mineral claims that carry the right of sale or licensing would be completely satisfactory.

It must be emphasized that no entity that has met the basic requirements of either human presence (NASA astronauts on the Moon), or *in situ* representation by an unmanned spacecraft (American Viking and Pathfinder and Soviet Mars Landers on the Martian surface), has ever claimed property rights or national sovereignty. Certainly, those individuals who presently purport to be selling tracts of land on the Moon and Mars have no basis whatsoever for claiming ownership in the first place. They lack both the right to sell what is not theirs and the legal or juridical authority to act as registrars of ownership. The existence of such schemes serves to discredit legitimate mineral rights or ownership claims based on reasonable criteria of presence. Similarly, the existence of ad hoc registries for claims, such as that of Professor Lawrence D. Roberts of the Archimedes Institute, can be viewed in the same light as the proceedings of a moot court. The existence of this registry may serve to stimulate awareness of the problem, so that some official mechanism for claim registry can be brought into existence. The issue of claim registration is made timely and urgent by the impending Near-Earth Asteroid Prospector (NEAP) mission of the Space Development Corporation. The time for initiating a recognized claim registry system has arrived.

BIBLIOGRAPHY

1. McKay, M.F., D.S. McKay, and M.B. Duke. Space Resources, NASA SP-509, Washington, DC, 1992.
2. Lewis, J.S., M.S. Matthews, and M.L. Guerrieri (eds). *Resources of Near-Earth Space*. University of Arizona Press, Tucson, 1993.
3. Lewis, J.S. *Mining the Sky*. Addison-Wesley, Reading, MA, 1996.
4. Heiken, G., D. Vaniman, and B. French. *Lunar Sourcebook*. Cambridge University Press, Cambridge, 1991.
5. Taylor, L.A., and W.D. Carrier III. Oxygen production on the Moon: An overview and evaluation. In J.S. Lewis, M.S. Matthews, and M.L. Guerrieri (eds), *Resources of Near-Earth Space*. University of Arizona Press, Tucson, 1993.
6. Wittenberg, J.L., J.F. Santarius, and G.L. Kulcinski, Lunar Source of ^3He for commercial fusion. *Fusion Tech.* 10: 167–178 (1986).
7. Boston, P.J. (ed.). *The Case for Mars*. Univelt, San Diego, 1984.
8. McKay, C.P. (ed.). *The Case for Mars II*. Univelt, San Diego, 1985.
9. Stoker, C.R. (ed.). *The Case for Mars III*. Univelt, San Diego, 1987.
10. Meyer, T.R. (ed.). *The Case for Mars IV*. Univelt, San Diego, 1990.
11. Schmidt, S., and R. Zubrin (eds). *Islands in the Sky*. Wiley, New York, 1996.
12. Melosh, H.J., and A.M. Vickery. Impact erosion of the primordial Martian atmosphere. *Nature* 338: 487–489 (1989).
13. Ash, R.L., W.L. Dowler, and G. Varsi. Feasibility of rocket propellant production on Mars. *Acta Astronautica* 5: 705–724 (1978).

14. Hepp, A.F., G.A. Landis, and C.P. Kubiak. A chemical approach to carbon dioxide utilization on Mars. In J.S. Lewis, M.S. Matthews, and M.L. Guerrieri (eds), *Resources of Near-Earth Space*. University of Arizona Press, Tucson, 1993, pp. 799–818.
15. Meyer, T.R., and C.P. McKay. The atmosphere of Mars—resources for the exploration and settlement of Mars. In P.J. Boston (ed.), *The Case for Mars*. Univelt, San Diego, 1984, pp. 209–232.
16. Farmer, C.B., and P.E. Doms. Global seasonal variation of water vapor on Mars and the implications for permafrost. *J. Geophys. Res.* 84: 2881–2888 (1979).
17. McKay, C.P., T.R. Meyer, P.J. Boston, M. Nelson, T. MacCallum, and O. Gwynne. Utilizing Martian resources for life support. In J.S. Lewis, M.S. Matthews, and M.L. Guerrieri (eds), *Resources of Near-Earth Space*. University of Arizona Press, Tucson, 1993, pp. 819–843.
18. Burns, R.G. Gossans on Mars. *Lunar Planet. Sci.* XVIII: 713–721 (1988).
19. Gehrels, T. (ed.). *Hazards Due to Asteroids and Comets*. University of Arizona Press, Tucson, 1994.
20. Lewis, J.S. *Rain of Iron and Ice*. Addison-Wesley, Reading, MA, 1996.
21. Lewis, J.S. *Comet and Asteroid Impact Hazards on a Populated Earth: Computer Simulations*. Academic Press, San Diego, 1999.
22. Glaser, P.E., F.P. Davidson, and K. Csigi (eds). *Solar Power Satellites: A Space Energy System for Earth*. Wiley-Praxis, Chichester, UK, 1998.
23. Criswell, D.R. Solar power system based on the Moon. In P.E. Glaser (ed.), *Solar Power Satellites: A Space Energy System for Earth*. Wiley-Praxis, Chichester, UK, 1998, pp. 599–621.
24. Feingold, H., M. Stancati, A. Friedlander, M. Jacobs, D. Comstock, C. Christensen, G. Maryniak, and J.C. Mankins. *Space Solar Power: A Fresh Look at the Feasibility of Generating Solar Power in Space for Use on Earth*. SAIC-97/1005, 1997.
25. *Agreement Governing the Activities of States on the Moon and Other Celestial Bodies*. U.S. Government Printing Office, Washington DC, 1980.

JOHN S. LEWIS
University of Arizona
Tucson, Arizona

SPACE SHUTTLE ORBITER

Introduction

The first serious mention of a fully reusable launch vehicle in a widely circulated publication was in 1952. In that year, Collier's Magazine ran a series of articles in which the concept of a fully reusable space launch vehicle was developed to have a vehicle that would transport cargo and people from the ground to an orbiting space station. The principal author of this article was Dr. Wernher von Braun (1).

It took 10 years from the publication of the article in Collier's before the idea of a reusable space ship was seriously considered. During the 1960s, the U.S. Air Force performed detailed studies of a reusable space ship called "Dynasoar," which was intended as a manned reconnaissance vehicle. NASA also conducted

14. Hepp, A.F., G.A. Landis, and C.P. Kubiak. A chemical approach to carbon dioxide utilization on Mars. In J.S. Lewis, M.S. Matthews, and M.L. Guerrieri (eds), *Resources of Near-Earth Space*. University of Arizona Press, Tucson, 1993, pp. 799–818.
15. Meyer, T.R., and C.P. McKay. The atmosphere of Mars—resources for the exploration and settlement of Mars. In P.J. Boston (ed.), *The Case for Mars*. Univelt, San Diego, 1984, pp. 209–232.
16. Farmer, C.B., and P.E. Doms. Global seasonal variation of water vapor on Mars and the implications for permafrost. *J. Geophys. Res.* 84: 2881–2888 (1979).
17. McKay, C.P., T.R. Meyer, P.J. Boston, M. Nelson, T. MacCallum, and O. Gwynne. Utilizing Martian resources for life support. In J.S. Lewis, M.S. Matthews, and M.L. Guerrieri (eds), *Resources of Near-Earth Space*. University of Arizona Press, Tucson, 1993, pp. 819–843.
18. Burns, R.G. Gossans on Mars. *Lunar Planet. Sci.* XVIII: 713–721 (1988).
19. Gehrels, T. (ed.). *Hazards Due to Asteroids and Comets*. University of Arizona Press, Tucson, 1994.
20. Lewis, J.S. *Rain of Iron and Ice*. Addison-Wesley, Reading, MA, 1996.
21. Lewis, J.S. *Comet and Asteroid Impact Hazards on a Populated Earth: Computer Simulations*. Academic Press, San Diego, 1999.
22. Glaser, P.E., F.P. Davidson, and K. Csigi (eds). *Solar Power Satellites: A Space Energy System for Earth*. Wiley-Praxis, Chichester, UK, 1998.
23. Criswell, D.R. Solar power system based on the Moon. In P.E. Glaser (ed.), *Solar Power Satellites: A Space Energy System for Earth*. Wiley-Praxis, Chichester, UK, 1998, pp. 599–621.
24. Feingold, H., M. Stancati, A. Friedlander, M. Jacobs, D. Comstock, C. Christensen, G. Maryniak, and J.C. Mankins. *Space Solar Power: A Fresh Look at the Feasibility of Generating Solar Power in Space for Use on Earth*. SAIC-97/1005, 1997.
25. *Agreement Governing the Activities of States on the Moon and Other Celestial Bodies*. U.S. Government Printing Office, Washington DC, 1980.

JOHN S. LEWIS
University of Arizona
Tucson, Arizona

SPACE SHUTTLE ORBITER

Introduction

The first serious mention of a fully reusable launch vehicle in a widely circulated publication was in 1952. In that year, Collier's Magazine ran a series of articles in which the concept of a fully reusable space launch vehicle was developed to have a vehicle that would transport cargo and people from the ground to an orbiting space station. The principal author of this article was Dr. Wernher von Braun (1).

It took 10 years from the publication of the article in Collier's before the idea of a reusable space ship was seriously considered. During the 1960s, the U.S. Air Force performed detailed studies of a reusable space ship called "Dynasoar," which was intended as a manned reconnaissance vehicle. NASA also conducted

studies of reusable space ships. Following a thorough review, a decision was reached in 1969 to assign all manned space operations to NASA. Accordingly, the Air Force's "Dynasoar" was cancelled.

Early in 1969, the Management Council of NASA's Office of Manned Space Flight, chaired by Associate Administrator George Mueller, met several times. Dr. Wernher von Braun, the director of NASA's George C. Marshall Space Flight Center and a member of the Management Council, vigorously advocated adopting the idea of a fully reusable space ship that he had written about in *Collier's Magazine* 17 years earlier. Eventually, a consensus was reached that a fully reusable space ship should be developed.

Dr. Robert Gilruth, the director of the NASA Manned Spacecraft Center in Houston, Texas (now the NASA-Johnson Space Center), was also a member of the Management Council. When he returned to the Center, he asked Dr. Max Faget, his Director of Engineering and Development (E&D) to study the problem of creating a reusable space ship. The essential problem was that the vehicle had to be able to fly both in the atmosphere as well as in space. Dr. Faget considered the most difficult problem that of returning the vehicle from orbit. He was the key individual in developing the successful Earth atmospheric entry techniques for the Mercury, Gemini, and Apollo spacecraft. Thus, Dr. Faget again looked to the high angle of attack and blunt body as the solution of the problem. The entry heating could be concentrated on the bottom of the vehicle thereby minimizing the weight of the thermal protection required. Using a handmade balsa model he built, Faget demonstrated that such an airplane had a stable, extremely high angle of attack that made the concept feasible. Range could be controlled by rolling around the velocity vector as was done by previous manned spacecraft. Faget believed that a straight wing rather than a delta wing would provide better subsonic landing performance.

While the studies were being performed to see what the Shuttle systems would look like, estimates of the funding were being established that would have a great effect on the design. Economic studies (2) were also being conducted to show the advantage of a reusable launch system. It was determined that the development cost of the fully reusable Shuttle was too high; therefore, the fully reusable booster was eliminated, and the increased operational cost was accepted (3).

Preliminary Design Considerations

General Outline. Technology studies conducted before initiating the Shuttle design indicated that three principal developments would be necessary to achieve the desired performance of the Shuttle. It was determined that if problems occurred, the development risk and possible cost increases were worth the enhanced performance. The three developments were the dual-cycle main rocket engine, the reusable surface insulating (RSI) thermal protection system for the orbiter, and an innovative flight control system. If development problems occurred, it was considered that the budget would allow recovery if only three new systems were designed. Existing technology or only small incremental improvements would be employed for the remaining components of the system.

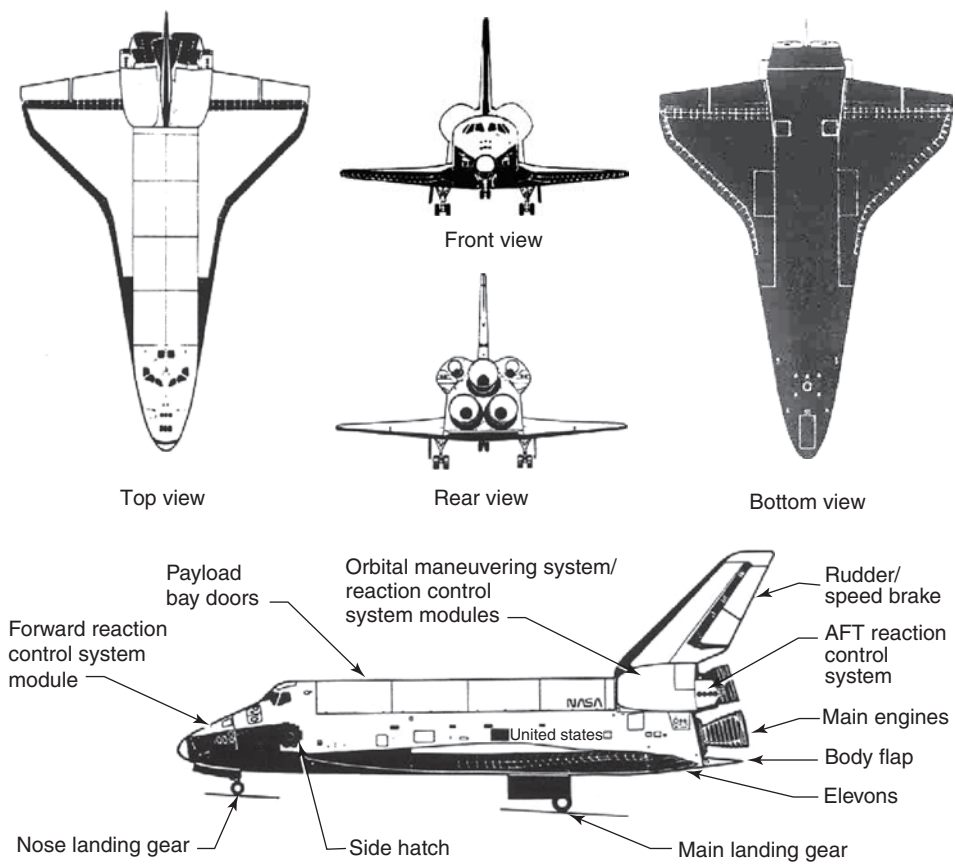
Phase A studies were conducted to determine basic requirements and their effect on the design. The principal issues were the size and weight of the payload and the cross range requirement for the orbiter. The size and weight of the payload were determined by the requirements of reconnaissance satellites that would be launched using the shuttle. This fixed the payload at 65,000 pounds at takeoff and a 35,000 pound landing capability. The size and shape of the payload bay were set at 60 feet in length and 15 feet in diameter to accommodate the largest national security related payloads. The cross range had to be 1000 miles to provide what is called a "once around" capability for manned reconnaissance missions. This means that the shuttle could execute one polar orbit and return to the original launch site because Earth rotates a distance of about 1000 miles at the equator in the time it takes to execute one orbit.

Specific details of design approaches were studied. Heat-resistant structures were compared to a reusable surface insulating thermal protection system, a hypergolic reaction control system was compared to a more advanced liquid hydrogen-liquid oxygen system, and a fly by-wire flight control system design were the subjects of some of the typical studies. Wind tunnel tests were conducted to determine wing size and configuration. Air breathing jet engines were initially proposed for the flyback capability of both the booster and the orbiter but it was determined that they would be too heavy for the performance gain. Entry technique, cross range requirement, landing speed, and the approach pattern had to be designed. A method to transport the Orbiter needed to be studied to move it from some of the landing sites to the launch pad.

Phase A study contracts were awarded to three competitive teams: Rockwell, North American, and General Dynamics, Convair; Martin Marietta, and Mc Donnell; and Boeing and Lockheed for preliminary designs. Upon completion of Phase A, a competition was held for the Phase B detailed design of the vehicle. The Rockwell, North American-General Dynamics, Convair and the Martin Marietta-Mc Donnell teams were selected to design both a high and low cross range configuration. The Phase A and Phase B studies led to the design of a fully recoverable orbiter, a disposable fuel tank, and parachute-recoverable solid rocket boosters. High performance hydrogen-oxygen engines were placed in the Orbiter to recover these high cost units after each flight.

Slightly later, a study was also awarded to Grumman and Boeing for more innovative designs, and a further study contract was given to Lockheed which had proposed a novel "stage and a half" concept. The studies were extended because budget concerns had not been settled, and a decision between fully reusable system and the development of only one fully reusable vehicle had not been made. More than 144 different configurations were examined; 41 were wind tunnel tested before the final design was accepted. The Air Force cross range requirement dictated the delta wing configuration for the orbiter shown in Fig. 1.

The final configuration consisted of the components shown in Fig. 2, the delta wing orbiter, an externally carried fuel tank for the liquid hydrogen and liquid oxygen that run the Space Shuttle Main Engine (SSME), and two solid rocket boosters attached to the tank. The three Space Shuttle Main Engines and the two solid-fueled boosters are powered up on liftoff, and the solid rocket boosters are jettisoned when they burn out in about 2.5 minutes. The empty



Dimensions and weight

Wing span	23.79 m	(78.06 ft)
Length	37.24 m	(122.17 ft)
Height	17.25 m	(56.58 ft)
Tread width	6.91 m	(22.67 ft)
Gross takeoff weight		Variable
Gross landing weight		Variable
Inert weight (approx)	74 844 kg	(165 000 lb)

Minimum ground clearances

Body flap (AFT end)	3.68 m	(12.07 ft)
Main gear (door)	0.87 m	(2.85 ft)
Nose gear (door)	0.90 m	(2.95 ft)
Wingtip	3.63 m	(11.92 ft)

Figure 1. A detailed representation of the reusable Orbiter. The locations of the three Space Shuttle Main Engines (SSME) and the Orbital Maneuvering System (OMS) engines are clearly shown. The control surfaces are also delineated. The figure is taken from NASA website: <http://spaceflight.nasa.gov/history/shuttle-ir/multimedia/diagrams/shuttle/shuttle-1.htm>.

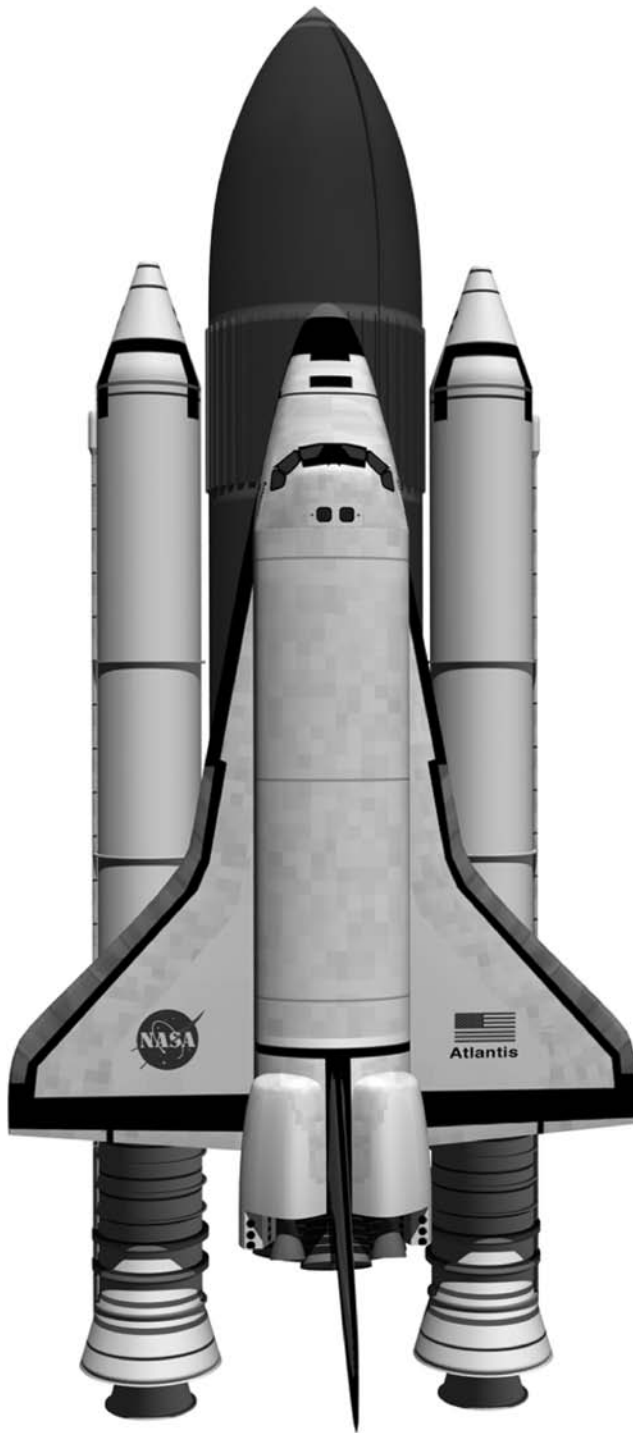


Figure 2. A drawing of the final configuration of the Space Shuttle viewed from the top. It shows the relative locations of the orbiter, the fuel tank, and the two solid-fueled boosters. The figure is taken from the following NASA website: <http://www.hq.nasa.gov/office/codeq/risk/workshop/bover.doc>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

rocket cases are parachuted into the ocean where they are recovered and towed back to the launch site. The main engines keep running for six or seven minutes until the fuel is exhausted. The external tank is then jettisoned and is burned up on reentry into the atmosphere. The Orbiter then goes into Earth orbit using the orbital maneuvering engines. As already stated, the Orbiter reenters the atmosphere, when the mission is finished, at a large angle of attack—more than 40°—to dissipate the orbital energy. The parameters of the system are shown in Table 1.

Table 1. **Orbiter Specifications**

	<i>Dimensions</i>	
Total length	37.24 meters	122.17 feet
Height	17.25 meters	56.58 feet
Wingspan	23.79 meters	78.06 feet
Vertical stabilizer	8.0 meters	26.31 feet
Tread width	6.9 meters	22.67 feet
Wing length	18.3 meters	60 feet
Wing max thickness	1.5 meters	5 feet
Eleven inboard	4.2 meters	13.8 feet
Eleven outboard	3.8 meters	12.4 feet
Payload bay length	18.3 meters	60 feet
Payload bay diameter	4.6 meters	15 feet
Payload volume	148.6 square meters	1600 square feet
Crew cabin	71.5 cubic meters	2525 cubic feet
	<i>Weight</i>	
Empty	74,844 kilo	165,000 lb
Landing with payload	96,163 kilo	212,000 lb
	<i>Thrust</i>	
Orbiter, sea level	1,668,000 newtons	375,000 lb each engine
	<i>Orbital maneuvering system</i>	
Thrust vacuum	26,688 newtons	6000 lb
ISP	313 seconds	
Chamber pressure	125 psi	
Mixture ratio	1.65:1	
Propellant	10,830 kilo	23,876 lb
	<i>Reaction control system</i>	
Primary 38, forward 14, aft 2 each pod		
Thrust, vacuum	3870 newtons	870 lb
Vernier 6, forward 2, aft 4		
Thrust vacuum	106 newtons	24 lb

Table 1. (Continued)

<i>Electrical power</i>		
Voltage	28 Volts	
Power	2 kw at 32.5 Vdc, 61.5 amps 12 kw at 27.5 Vdc, 436 amps	
<i>Atmospheric revitalization</i>		
Pressure	760 mm Hg \pm 103	14.7 \pm 2 psi
Ratio	79% N ₂ , 21% O ₂	
<i>Auxiliary power unit</i>		
Power	135 hp	
Weight	39 kg	88 lb
Fuel	158 kg	350 lb

Engineering Management Considerations. The Space Shuttle was entirely different from any air/spacecraft that had ever been designed and built. The engineering teams in the NASA Office of Manned Space Flight and the associated institutions had to be organized properly. It was decided to develop in-house engineering design groups. At the NASA Johnson Space Center, key engineers were assigned to a group housed in Building 36 to design a “DC-3” (the first successful passenger airliner) space transportation vehicle. The group soon found the task of designing control systems, propulsion systems, structure, and hydraulic systems more challenging than initially thought, and so more time was required before a complete vehicle design was attempted. This effort, however, was excellent training for the government team that would later lead to a NASA supervised contractor to design and manufacture the actual flight vehicles. Thus, NASA acted as the system integrator for the Space Shuttle.

After a short extension of the Phase B studies, the competition for the Orbiter hardware contract was held, and the contract was awarded to Rockwell, North American in 1972. Later, the external tank hardware contract was awarded to Martin Marietta, and the solid rocket motor boosters to Thiokol, Inc.

The Space Shuttle Main Engine

The principles of rocket propulsion are described elsewhere in this Encyclopedia (see Liquid-Fueled Rockets). The reusable SSME was a major advance in rocket engineering. To propel the Space Shuttle into Earth orbit, it had to have a higher thrust-to-weight ratio than any previously high-thrust rocket engine. The higher thrust-to-weight ratio is achieved by operating the SSME at a chamber pressure

higher than that of any previous rocket. The high chamber pressure (2960 psi) is achieved by using part of the liquid hydrogen–liquid oxygen supply to operate two high-pressure turbopumps, one for hydrogen and another one for oxygen, that provide the fuel and oxidizer to the combustion chamber at a very high flow rate. The nozzle of the main engine is cooled by liquid hydrogen (regenerative cooling) before it is fed to the combustion chamber and burned. The arrangement described here is called a two-stage or dual-combustion system, and the flow diagram for the SSME is shown in Fig. 3. The thrust developed by the SSME, as finally built, is 375,000 lbs.

The concept of the dual-combustion engine was initiated by the Air Force, which had Pratt & Whitney develop and test the high-pressure pumps for the dual-cycle rocket engine. These pumps were built for an XRL-129 engine at the 25,000-pound thrust level.

A competition was held for the contract, and proposals were received from Aerojet, Pratt & Whitney, and Rocketdyne. Rocketdyne was selected to build the engine for 375,000-pound thrust at sea level. The contract, initiated in 1970, preceded the award of the Orbiter contract because it was believed that more time was necessary to develop the engine (4).

The entire space shuttle propulsion system has a thrust of about 7 million pounds, about 375,000 pounds for each of the three SSMEs on the Orbiter, and

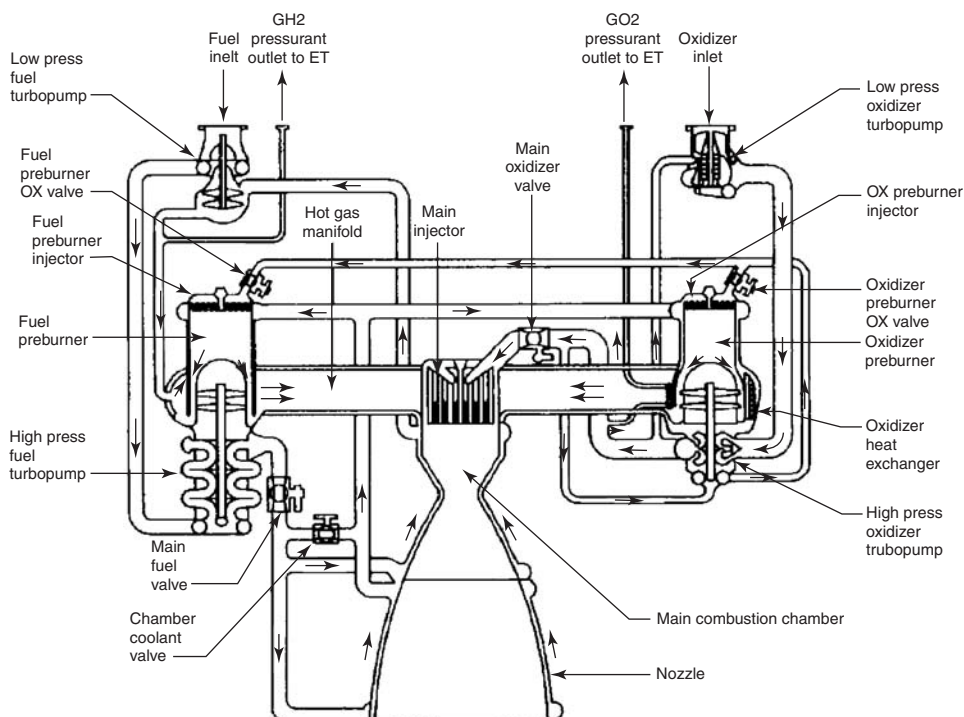


Figure 3. The main engine schematic. A drawing of the flow diagram of the fuel and the oxidizer in the Space Shuttle Main Engine. The figure is taken from NASA website: <http://www.shuttlepresskit.com/scom/216.pdf>.

2.9 million pounds for each of the solid rocket boosters. The SSMEs can be throttled from 65 to 109% of their rated power level. This thrust is enough to lift the whole space shuttle stack, according to the figures shown in Table 1.

Orbiter System Descriptions

Crew Compartment. The crew compartment has provisions for seven members. Four are on the flight deck, and three are on the mid-deck (Fig. 4). The commander uses the left seat and the pilot the right. Behind the pilot sits the mission specialists, and in the center rear of the flight deck sits the payload specialist. Three additional mission specialists can be seated on the mid-deck. Flight controls and displays are in the usual aircraft positions at the front of the flight deck and a mission or systems panel is located on the right side. The rear of the flight deck has windows that look out into the payload bay and above. These windows also are used to view the manipulation of the payloads. The mid-deck is used for habitability and contains sleep stations, a galley, and a waste collection system. An air lock is located in the middle of the aft portion for external vehicle access through the cargo bay. The forward portion of the mid-deck is used for storage lockers, and behind the lockers are the flight computers and other avionic equipment.

During a space mission, the crew compartment is operated in a “shirt-sleeve” environment with an atmosphere of oxygen and nitrogen at about 14.7 psi at room temperature (300 K). The air regeneration and waste elimination

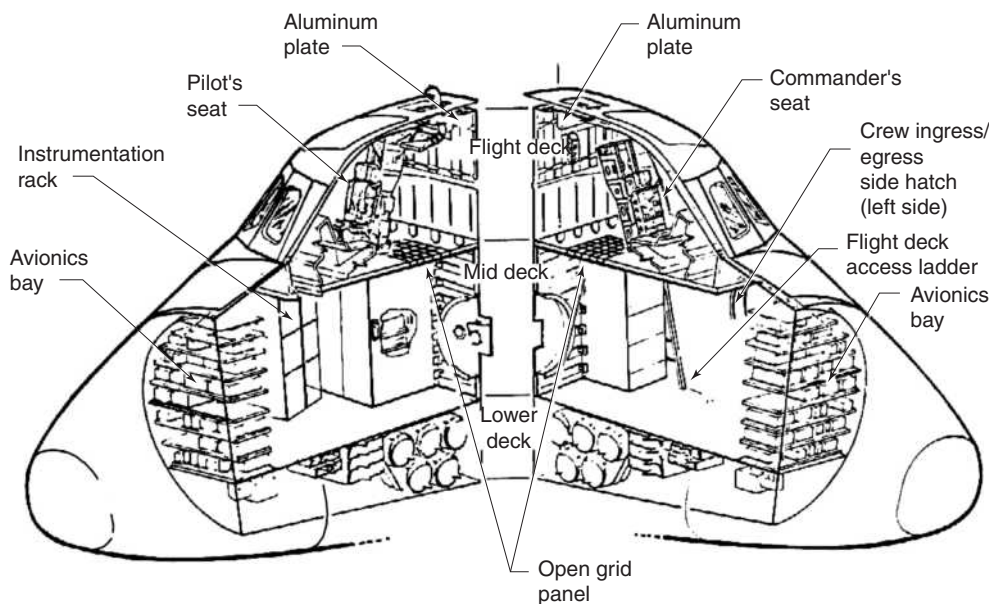


Figure 4. A cutaway drawing of the crew compartment as well as some detailed views of the structure. The figure is taken from the NASA website: <http://spaceflight.nasa.gov/history/shuttle-ir/multimedia/diagrams/shuttle/shuttle-6.htm>.

systems are part of the crew compartment. The flight duration for a given crew size is determined by the power available and the exhaustion of other consumables in the crew compartment. For an “average” mission, the Space Shuttle Orbiter can remain in Earth orbit for about 2 weeks.

Guidance, Navigation, and Control (GN&C). The Space Shuttle Orbiter was one of the first operational applications of digital fly-by-wire systems for aircraft flight control. The design expanded on the experimental work conducted by the NASA Dryden Flight Research Center with a modified F-8 aircraft. The GN&C system responds to software commands to provide vehicle control and provides information to the sensors and control data to the flight computers that control the flight. The Orbiter’s five computers are arranged into a redundant set of four that comprise the primary set; the fifth computer is used as independent, backup, flight control system. Data from vehicle sensors are transmitted through a multiplexer/demultiplexer (MDM) by data bus wire (there are no direct mechanical linkages to the various controls) to the computers, which, using the MDM, send commands to the vehicle control effectors. The system consists of two modes of operation, “auto,” where the computers do all the flying, and “control stick” steering, where the pilot can manually introduce commands into the computer system. Multifunctional cathode ray tubes (CRT) display the system’s status and flight information to the crew.

The navigational interfaces feed data to the general purpose computer (GPC) to do the navigation and provide the guidance information to the system to drive the control effectors. Air data probes are deployed at subsonic velocity at the end of the flight. A Tacan system was also used to navigate during final flight. Later, the vehicle was modified to accept Global Precision System (GPS) inputs. The backup flight control system has a single-string system that uses some different sensors and was programmed differently from the redundant set. The GPC is also used to operate other vehicle systems.

Hydraulic System. Hydraulic power is provided for the aerodynamic flight control system, main engine gimbaling and control, nose wheel steering and brakes, and other items. The power source is a hydrazine-fueled auxiliary power unit (APU). Three redundant independent systems can provide full power using only two APUs and reduced rate power using a single system. The systems are active during launch and entry to provide for the engine controls during launch, flight controls during launch, and flight controls during entry and landing approach. The hydraulic pressure supplied is 3000 psi. Deployment of the landing gear is by gravity force only; however, hydraulic actuators are used to stop the free-fall gear because the stopping loads would otherwise be at the design load levels for the entire wing center section.

Auxiliary Power Unit. The auxiliary power unit (APU) produces power for the Orbiter’s hydraulic system. Three separate systems are used to power hydraulic actuators for flight control during boost and entry. The three APUs are located in the aft fuselage. The APU uses catalytic decomposition of the hydrazine fuel (N_2H_4) and creates hot gas to drive a two-stage turbine, which drives a hydraulic pump. The rated power of each APU is 135 horsepower. The APU was built by the Sundstrand Corporation, Rockford, Illinois.

Orbital Maneuvering System (OMS). The OMS provides the thrust for orbital insertion, orbital circulation, orbital transfer, rendezvous, and deorbit. Two

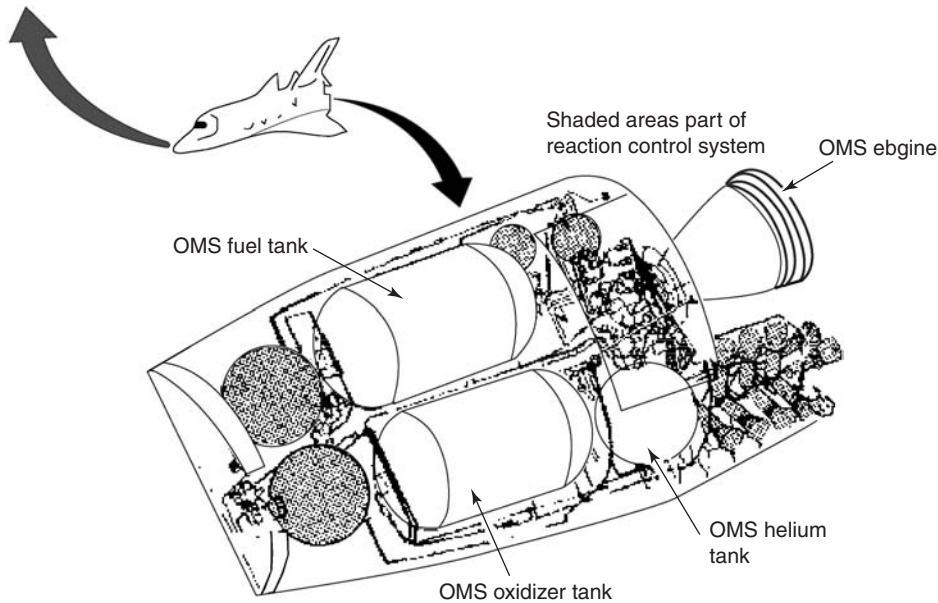


Figure 5. A diagram of the OMS and the RCS systems. These systems are closely linked. The figure is taken from the website: <http://faculty.erau.edu/ericksol/shuttle/steve'sproject/m1/s1-10aoms.html>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Pods (Fig. 5) on the aft fuselage house the two 6000-pound-thrust OMS engine systems, as well as the aft reaction control system (RCS), which is redundant and can be used if the OMS engine fails. The OMS is pressurized by helium and uses nitrogen tetroxide (N_2O_4) and monomethyl hydrazine (MMH), Earth storable, hypergolic propellants. The engine nozzle extension is fabricated from aluminum alloy and is cooled by radiation. The combustion chamber walls are regeneratively cooled by the fuel before it is fed to the engine injector. Two electromechanical gimbal actuators control each engine. The pods are cross-fed, which allows using propellant from either pod. The OMS pod was built by the McDonnell Douglas Astronautics Company of St. Louis, Missouri. The engine was built by the Aerojet Liquid Rocket Company, Sacramento, California.

Reaction Control System. The RCS (Fig. 5) provides attitude control in pitch, roll, and yaw and small velocity changes in translation above 70,000 feet altitude. The RCS thrusters are located in the forward nose area and in both OMS/RCS aft pods. The system is pressure fed and uses the same propellants as the OMS, which can be cross-fed so that the fuel from the OMS tanks can be used. The forward RCS has 14 primary engines of 870 pounds thrust each, and two vernier engines of 24 pounds thrust each. Each aft pod has 12 primary and two vernier engines. The RCS engines were built by the Marquardt Company, Van Nuys, California.

Communications. An S-band (2–4 GHz) communication system is the primary means of communicating between the Orbiter, and ground stations during the ascent, orbital, and entry phases of flight (Fig. 6). There is one uplink from

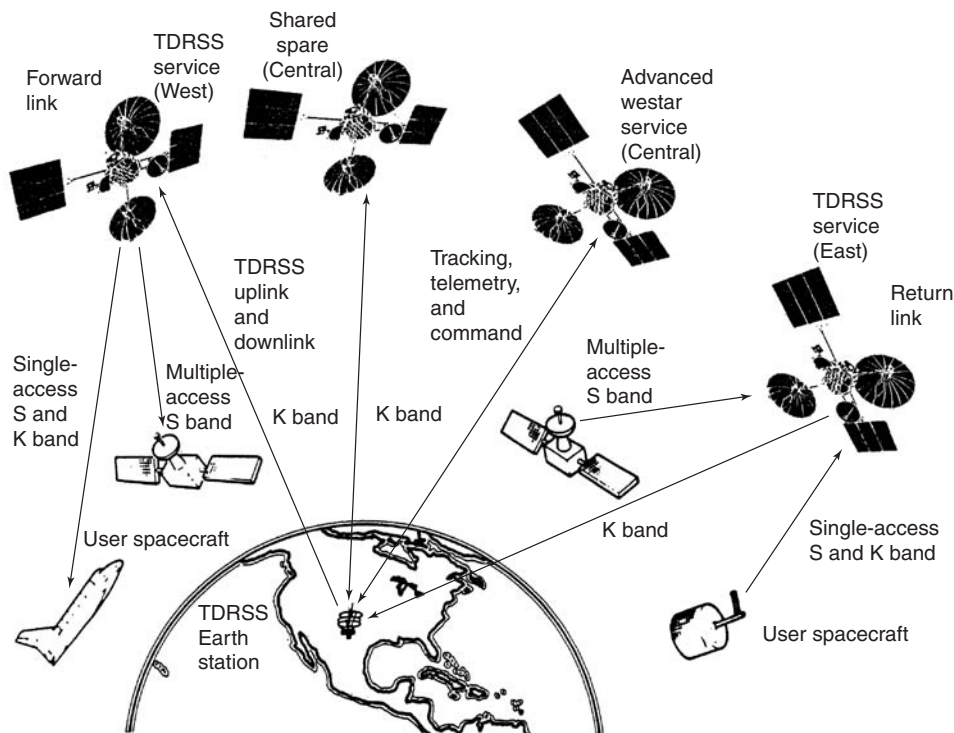


Figure 6. Communication System linking four identical and interchangeable satellites with earth station (diagram courtesy of NASA).

the ground which is phase-modulated (PM) and provides commands, voice, and coherent signals to the Orbiter. There are three downlinks; one is phase-modulated and provides voice, telemetry, two-way Doppler, and a coherent ranging signal. The other two are frequency-modulated (FM) for real-time data and video.

In addition to the S band, an ultra-high-frequency (UHF) band is used for air to air and air to ground voice, and the Ku band is used with the Tracking and Data Relay Satellite (TDRS). There are four tracking and data relay satellites in geosynchronous orbit, which provide continuous real-time communications with the ground control station. UHF is used during landing for aircraft communications and during extravehicular activities (EVA). The Ku band provides a much higher gain signal for higher data rates. The system also includes a rendezvous radar.

Electrical Power. Three fuel cells generate 28-volt dc electrical power through a chemical reaction. The fuel cells use hydrogen and oxygen that is stored in a supercritical condition. The dc power is routed to a three-bus system that distributes the power. The fuel cells were built by Pratt & Whitney, East Hartford, Connecticut.

Environmental Control and Life Support System (ECLSS). The Atmospheric Revitalization Control System (ARCS) maintains a habitable

environment for the crew and passengers and a conditioned environment for avionic equipment inside the crew cabin. The crew cabin is pressurized to 14.7 psia with 79% nitrogen and 21% oxygen. Oxygen is obtained from a super-critical cryogenic storage supply and can also use the oxygen supply system of the fuel cells.

A water coolant loop subsystem (WCLS) is used to condition the crew cabin thermally by collecting heat through air-to-water heat exchangers and transferring heat to the Freon[®] coolant loop.

Payload Deployment and Retrieval System. The Remote Manipulator System (RMS) is the mechanical payload deployment and retrieval system used to release and capture payloads. The basic system consists of a manipulator arm, display and control panel, and interface units, that interface with the Orbiter's computers. The arm is located on the left (port) side longeron of the orbiter's payload bay. It is 50 feet, 3 inches long, 15 inches in diameter and has six degrees of freedom. The booms are made of carbon composite material, and the joints are of aluminum alloy. It can use standard or special purpose end effectors. The RMS has an active and passive thermal control system. A jettison system is available if the RMS cannot be stowed. The system was supplied by Spar Aerospace of Canada.

Thermal Protection System. Because of its importance in total vehicle weight, one of the most studied systems for the Orbiter was the TPS. The Orbiter was intended as a reusable vehicle, so that the ablative thermal protection used on previous people-carrying spacecraft was considered impractical. It was judged that an ablative system would have to be replaced on nearly every flight. The people at the NASA Research Center had spent much effort for a considerable time on reusable systems, particularly those that featured hot structures. Rene, coated columbium (niobium), and tantalum were some of the metals considered for various areas of the vehicle. Using the high angle of attack technique during the entry phase concentrates the heating on the bottom of the orbiter. The highest heating rates are experienced by the leading edges of the nose and wing. Different heat-resistant metals would be used for different regions of the Orbiter.

A different approach to thermal protection was being studied by people at the NASA Ames Research Center and by some contractors, Lockheed, General Electric, and McDonnell. The reusable surface insulator (RSI) was a new approach that shielded or insulated the heating from the vehicle by being a poor transmitter of the heat generated during atmospheric entry. This approach was considered favorably because the vehicle structure could be made of aluminum, with which the industry has much experience rather than less well understood high temperature metals. The RSI also somewhat uncoupled the TPS from the vehicle structure, which would allow easier adjustment if one or the other needed to be changed as flight experience matured the vehicle. Finally, using an RSI system allowed designing the structure before full understanding of the heating was obtained. Thus, it was possible for the Orbiter to perform the ALT program before the TPS design was complete.

The RSI is comprised of blocks, which were called "bricks" or "tiles" of about 6 × 6 inches. The material is a matrix of microfibers of high-purity silicon whose density is 8 pounds per cubic foot. The low density provides the poor conductivity

of the material. Because the material absorbs water easily, a thin borosilicate glass coating is applied to prevent the tile from absorbing water. Absorbed water would add considerable weight to the vehicle and would create other problems. The coefficient of expansion of the tile is much lower than that of the aluminum of the vehicle structure. Thus, the tile is bonded to the vehicle using a felt pad that accommodates the differential motion. A Room Temperature Vulcanizer (RTV) is used as the bonding agent.

Four types of thermal protection materials are used on the Space Shuttle air frame (Fig. 7).

- 1. High-temperature reusable surface insulation (HRSI) for areas of higher heating is used mostly on the bottom of the vehicle. These tiles are black.
- 2. Low-temperature reusable surface insulation (LRSI) for areas of lower heating is used mostly on the sides of the vehicle. These tiles are white.
- 3. Felt (coated nomex) reusable surface insulation (FRSI) for areas of low heating is used mostly on the top of the vehicle. This material is white.
- 4. Reinforced carbon-carbon (RCC) composite insulation for areas that experience the highest heating. This material is black.

The thickness of the tile is determined so that the bond-line temperature of the aluminum never exceeds 300°F, the temperature at which aluminum begins to lose its mechanical properties. After numerous cycles, the coating for the major heating area was black, whereas a coating for lower heating areas was white because of the different radiative emissivities of the materials. This was also

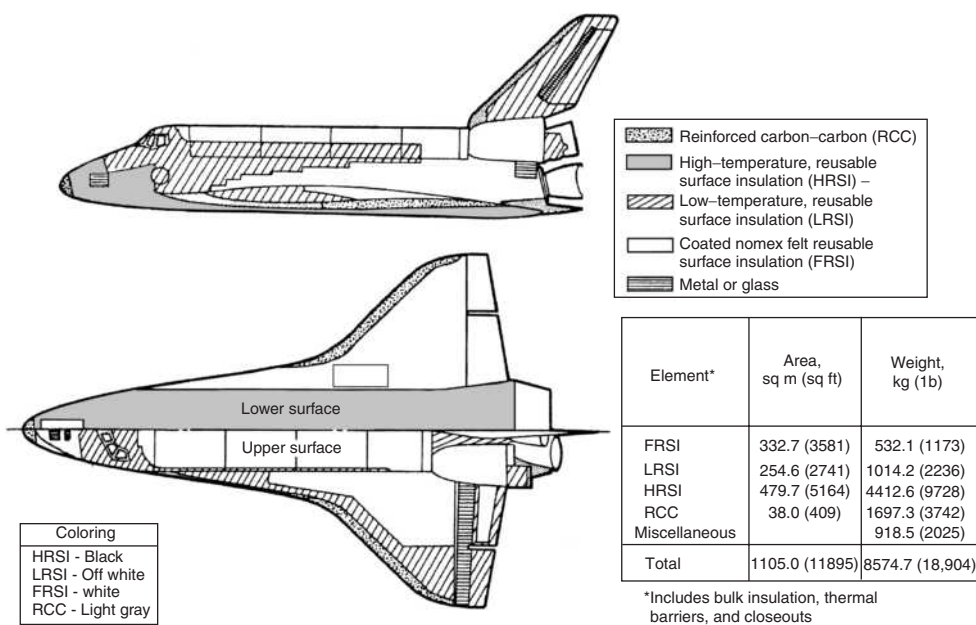


Figure 7. Thermal protection system, Orbiter 102 (diagram courtesy of NASA).

useful for orbit thermal control because of the different reflectivities of the materials when sunlight strikes them.

In areas such as the top of the vehicle, the felt pad used as a strain isolator in bonding the tile is coated with an RTV and is used on top of the wings and fuselage. Lockheed Sunnyvale was the supplier of the RSI.

Because of higher heating, a different approach was needed for the leading edges of the nose and wing. Carbon-carbon material was selected because it has the highest reusable temperature capability of all of the materials considered for the RSI. Of course, RCC is much denser than the other materials employed for the RSI system. Thus, it was necessary to minimize the area on which RCC was used. Great care was taken to perform accurate heating tests in developing the orbiter design to ensure that weight was minimized without endangering the vehicle when it reentered the atmosphere. This system was produced by LTV, Dallas, Texas.

Structure

The major structural components were made primarily from 2219 aluminum, a material well known to the aircraft industry. Due to the problem of an aft center of gravity, titanium was used in the aft main engine thrust structure because of its high strength to weight. This structure distributed the load from the three main engines into the Orbiter fuselage at the aft bulkhead of the payload bay. The payload bay was designed so that the payload doors did not take any bending loads. This simplified their design. The door material was changed to a graphite carbon composite as a major weight saving after an initial design. The crew cabin consisted of a double shell; the external shell takes the vehicle loads, and the inner shell sustained only cabin-pressure loads. This approach was considered a crew safety issue and was also used on the Apollo spacecraft.

Space Shuttle Test Flights

Approach and Landing Test (ALT) Program. The first Space Shuttle vehicle built was the “Enterprise”. The mold lines, the structure, and the weight distribution were essentially identical to the space capable vehicles that would follow. It would be used for flight tests in the atmosphere, “form fit and function” tests at the Kennedy Space Center launch site, and for public relations, such as the trip to the Paris Air Show in 1983 where the “Enterprise” was the centerpiece of the exhibitions.

The Approach and Landing Test Program was conceived as an early demonstration of the glide capability of the Orbiter in the lower atmosphere. It would test flight hardware if the orbital flight program was delayed. The method of selecting a reusable surface insulator (RSI) allowed designing the orbiter, while analysis continued to determine the thermal environment and to develop the parameters (thickness, weight, and composition) of the thermal protection system. The structure of the Orbiter and the thermal protection system were essentially uncoupled.

There was some concern for launching the orbiter from the top of a Boeing 747, which had been selected to ferry the Orbiter between the manufacturing site, a landing site, and the launch pad. Wind tunnel tests of the two vehicles were conducted, and a separation procedure was developed. It was discovered that the procedure used by Mayo, a British designer interested in ultra-long-range air transportation during the 1930s, which had a small seaplane (the "Mercury") carried by a flying boat (the "Mayo"), proved the best. This procedure put the upper aircraft in an attitude with a higher lift-to-drag ratio (L/D) than the carrier aircraft L/D, so that the Orbiter actually dropped the Boeing 747 aircraft when released.

The ALT program allowed demonstration of the innovative fly-by-wire flight control system. Only the Approach and Landing final flight phase software was required, so this also allowed a phased development process. A pilot-induced oscillation (PIO) problem was discovered in the software resulting from a priority limit technique on the hydraulic system. The ALT program offered the operations personnel the opportunity to exercise their systems on a new concept in manned spacecraft. The ALT program was initiated in 1977 and was successfully concluded after five Flights (6).

Space Flight-Test Program. "Columbia" was the first space shuttle to be completed. It flew for the first time on 12 April 1981 and returned to Edwards Air Force Base (AFB) on 14 April 1981. Two people were on board, Flight Commander John Young and Pilot Robert Crippen. Earlier first flights for manned spacecraft, Apollo, Gemini, and Mercury, were all performed without a crew. Thus, the risk incurred on previous first flights with a crew aboard was considerably smaller than it was for the Space Shuttle. Thus, special care had to be taken to ensure that nothing went wrong. (Note: The Soviet version of the Space Shuttle, "Buran," copied the American Space Shuttle's platform but did not carry reusable rocket engines. It flew only once without a crew, using an automated landing system. Because "Buran" did not have a redundant flight control system, the risk of putting people on board was deemed to be too high. "Buran" has never flown again (7).

The first four flights of "Columbia" were designated "test flights." During these missions, crews were limited to two people, and no experiments were on board. The purpose of these missions was to explore and to expand the flight envelope of "Columbia." Because of the risks involved, two special measures were taken. The most important was to provide ejection seats on the first flight and three subsequent flights. These seats were used previously on the SR-71 aircraft, and the Approach and Landing Test provided an escape capability for the crew during the first few minutes of flight.

A second concern was the possibility of a software failure of either the primary or the backup computer system. Thus, "manual" override provisions were made so that the crew could select a configuration of the computers that would work. Great care was also taken to select the most experienced and competent crews for the first four flights. The second flight of "Columbia" was executed in November 1981 with Richard Truly as commander and Joseph Engle as the pilot. Commander Jack Lousma and Pilot Gordon Fullerton flew the third flight in March 1982.

The final test flight was conducted in July 1982 with T. K. Mattingly as commander and Henry Hartsfield as pilot. Appropriately, "Columbia" landed at

Edwards AFB on 4 July 1982, and President Ronald Reagan was present to welcome the crew. Several hundred thousand people had driven out for the holiday to watch the spectacular landing (8).

Space Shuttle Operations

Space Shuttles have now flown more than 100 missions during the past 20 years. After the successful test flights by "Columbia" in 1981 and 1982, three other Orbiters were built and launched: "Discovery," "Challenger," and "Endeavor." Each was slightly different from the others, but by any measure, they were sister ships. The Space Shuttles have performed a wide variety of tasks that would never have been possible without them. They have launched a large number of commercial, military, and scientific satellites. They continue to perform these missions successfully, even though the commercial and military satellites have been removed from the Space Shuttle manifests for reasons of policy. The Shuttles have carried Spacelab into Earth orbit in which important scientific experiments have been performed. Finally, a number of Space Shuttle missions have been performed that have involved repairing, replenishing, and retrieving Earth orbiting satellites. The first of these flew in the Spring of 1984; during the flight, the disabled Solar Maximum (Solar Max) satellite was retrieved, repaired, and redeployed. Extravehicular activity (EVA) had to be employed in this mission, and it was considered quite dangerous at the time. Partly, as a result of the Solar Max mission, operations of this kind have become quite routine.

Correction to the Space Telescope Optics was another mission which demonstrated the feasibility of on-orbit repair of satellite repair.

On 28 January 1986, the Orbiter "Challenger" was lost in a tragic accident. It was replaced by the shuttle "Atlantis," and flights of the Space Shuttles resumed in September 1988, a little short of 3 years after the "Challenger" disaster. Because the "Challenger" disaster was not caused by a failure of a system in the Orbiter, the description of exactly what happened will be recounted in another article in this Encyclopedia.

Most recently, the Space Shuttle has been heavily involved in the construction and deployment of the components of the International Space Station. It is interesting that when the reusable space ship was first envisaged, transport back and forth from orbiting space stations was deemed the most important mission (1). There have been many modifications and changes in the Space Shuttle system since it was first fielded more than 20 years ago, but the basic configuration remains unchanged.

Many studies have been conducted to consider replacements for the Shuttle, but no system has yet shown enough promise to be built. The shuttle never accomplished its goal of substantially reducing the cost of delivering a payload to Earth orbit. There is some question whether the cost goals were ever realistic in the first place. Many reasons, such as flight frequency, operational considerations, and safety, can be put forth to explain why costs have remained high. Nevertheless, the Space Shuttle system has provided a safe method of continuing manned spaceflight. It is quite likely that the Space Shuttle will remain the United States' primary means for reaching Earth orbit in the foreseeable future.

BIBLIOGRAPHY

1. Collier's Magazine. Man will conquer space soon, March 22, 1952; Man on the Moon: The journey, October 18, 1952; Man on the Moon: The exploration, October 25, 1952; Man's survival in space: Testing the men, March 7, 1953; Man's survival in space: Emergency! March 14, 1953; The baby space station: The first step in the conquest of space, June 27, 1953. Can we get to Mars? Is there life on Mars?, April 30, 1954.
2. Heiss, K.P. Economic Analysis of the Space Shuttle System. Mathematica Corporation Report, January 31, 1972.
3. Heppenheimer, T.A. The Space Shuttle Decision: NASA's Search for a Reusable Space Vehicle. NASA SP-4221, 1999.
4. Covert, E. Technical Status of the Space Shuttle Main Engine. National Academy of Sciences Report, Washington, DC, February 1979.
5. Hallion, R.P., and J.O. Young. Space shuttle: Fulfillment of a dream. In R.P. Hallion The Hypersonic Revolution: Eight Case Studies in the History of Hypersonic Technology, Vol. II. Wright Patterson Air Force Base, Aeronautical Systems Division, Dayton, Ohio, 1987.
6. Logsdon, J. (ed.). Exploring the Unknown, Vol. IV, Accessing Space, NASA-SP-4407, 1999, pp. 173–174.
7. Newkirk, D. *Almanac of Soviet Manned Space Flight*. Gulf, Houston, 1990.
8. Yardley, J. STS-1 Postflight Mission Operation Report No. 989-81-01, May 12, 1981.

AARON COHEN
MILTON A. SILVEIRA
Space Shuttle Orbiter Project Office
Johnson Space Center
Houston, Texas

**SPACECRAFT GUIDANCE,
NAVIGATION AND
CONTROL SYSTEMS**

The difficulties in maneuvering a vehicle through space are not intuitively obvious to someone who has not been involved in the field. After all, we have been exposed to the television and movie concepts of space travel, wherein there seems to be almost unlimited electrical power, rocket thrust, and propellant, and the protagonist simply flies off to a distant moon or planet using visual references for navigation.

Actually, the finite limits of electrical energy, thrust, and propellant on a vehicle are the driving forces that demand efficient and accurate equipment to perform the functions of steering and navigation while keeping the vehicle attitude stabilized. Added to these restrictions are the needs for minimum volume and mass. An intricate system results that is extremely difficult to design, construct, and test and usually is one of the most expensive on the vehicle.

Terminology

As in most fields, the space industry has a set of terms and definitions peculiar to the field that, unfortunately, are not always consistent in meaning. The term “guidance system” is sometimes used interchangeably for “guidance and navigation system” or “guidance, navigation, and control system” or “guidance and control system.” In this article, the terms are used as follows:

Guidance means the actual steering of the vehicle as it travels through space. Guidance commands may originate from a crew onboard, from an onboard computer, or from external sources via radio commands. In addition, if the thrust of a space vehicle is variable and controllable, the command for modulating the thrust is usually a guidance function.

Navigation is the measurement of the location of the vehicle in space and plotting the course of the vehicle. Navigational fixes may come from onboard human sightings using telescopes and sextants, from automatic onboard star or horizon sensors, or from radio/radar tracking equipment on the ground.

Control refers to the spatial alignment and stabilization of the vehicle while the guidance and navigation functions are being performed, and includes onboard processing and routing of commands to the devices (typically thrusters, reaction wheels, control moment gyroscopes, or aerodynamic surfaces), termed *effectors*, that produce reactive forces on the vehicle.

The combination of these three functions into one system results in the *integrated guidance, navigation, and control system*, or simply the GN&C system.

Space Guidance

As mentioned earlier, the finite limits of electrical energy, thrust, and propellant on a space vehicle are the reasons one simply cannot point the vehicle toward the target and fly it there. Further, within the Universe, the target itself is typically moving, and if the spacecraft were simply steered toward it, the spacecraft could end up chasing it. The situation is analogous to that of a shotgunner who must “lead” the target, that is, aim at a point on the flight path ahead of the target so that the shot arrives at the point where the target will be rather than the point where it was when the trigger was pulled. Other complicating factors are the effects of the gravity of Earth, the Moon, the Sun, and the planets, as well as their individual rotations if the launch points or landing sites lie on the surface.

A typical spacecraft trajectory that results from these conditions has the following characteristics:

- The launch point is on a rotating Earth (or Moon or planet).
- The gravitational pull of Earth and the Moon are significant effects.
- The flight path is curved and is likely to be nonplanar.
- The flight duration may be anywhere from minutes to years.
- The flight path is carefully plotted and optimized during mission planning.

The guidance function steers the spacecraft along the preplanned flight path while accommodating these characteristics.

To understand the guidance function better, consider the following hypothetical mission trajectory:

1. Launch occurs at Cape Canaveral, Florida, U.S.A. The flight path bends from a vertical ascent to a due easterly direction. As orbital altitude, say 200 nautical miles, is achieved, the spacecraft flies on an elliptical path in a plane at the same angle to the equator as the latitude of the launch site—in this case, 28.5° .
2. The orbital altitude is then increased to 800 nautical miles by using the Hohmann transfer. This is a minimum energy maneuver where the thrusting is done at the apogee and perigee of the orbit.
3. A plane change is accomplished to permit rendezvous with another spacecraft in a different plane.
4. After rendezvous and docking with the second spacecraft, say, to replenish propellants, our spacecraft undocks and drifts away.
5. The spacecraft then thrusts in a new direction for transplanetary injection. Escape velocity is achieved, and the spacecraft leaves Earth on a hyperbolic trajectory toward the target planet.
6. Along the way, course corrections are made using small velocity increments (*delta V's* in space parlance).
7. The spacecraft flies to the target planet or passes nearby and continues out into space. These mission segments are shown pictorially in Fig. 1.

To accomplish the guidance function during each of these mission segments, the GN&C system must have reasonably precise knowledge of the pointing direction of the vehicle and the acceleration during thrusting. Devices used to obtain this information are called *inertial sensors* and include the gyroscope and accelerometer. Further, for the rendezvous phase, it needs to know the relative position of the spacecraft to the targeted vehicle; rendezvous radar can be used for this purpose. For the docking phase, a near-in distance-measuring sensor, such as a laser, is required.

Gyroscopes. The gyroscope, or gyro as it is usually called, is a spinning wheel supported in rings called gimbals. The gyro has two properties that are useful for guidance applications—the tendency to remain fixed in alignment in space if undisturbed, and the tendency to precess predictably when a torque is applied to the gimbal 90° to the spin axis.

Suppose that a gyro were mounted in a spacecraft. Suppose further that some method is used to spin the wheel and measure angular movement of the gimbals. A gyro used in this manner is termed a *vertical gyro*. Rotational motion of the spacecraft about all axes except the gyro spin axis could be measured and used for guidance purposes. For the case of spacecraft rotation about the gyro spin axis, a second gyro whose spin axis is at 90° to the spin axis of the first gyro could be added. Some aircraft and missiles that have short flight times use gyros in this manner for attitude control during flight (see, for example, the discussion in Ref. 1). However, drift due to bearing friction and difficulties in

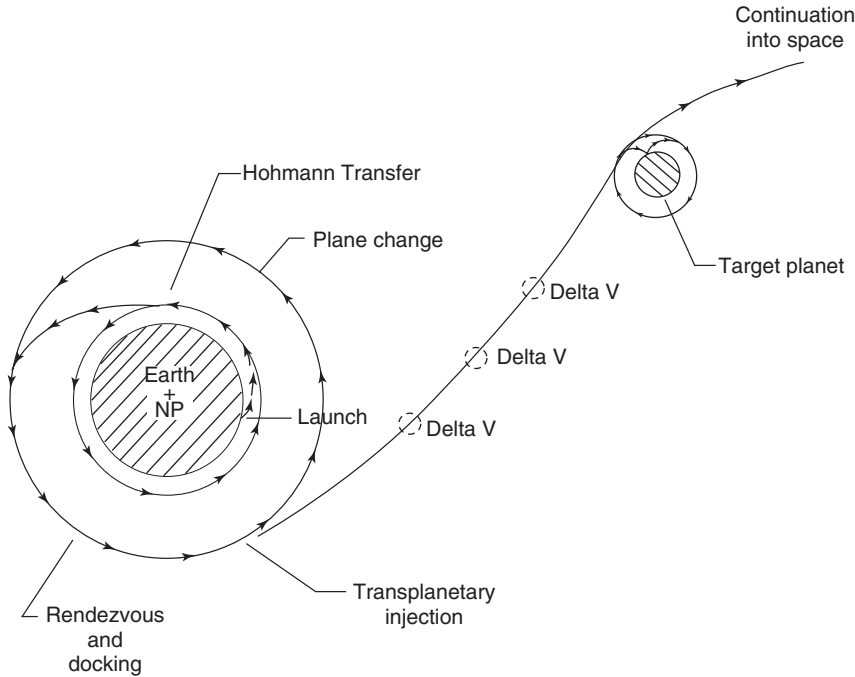


Figure 1. Hypothetical space mission.

precise readout of the gimbal positions usually make this application unsuitable for spacecraft that have precise trajectory requirements and long flight durations. In practical applications for modern spacecraft, the property of *precession* is more useful.

The spinning wheel in a gyro has an angular momentum and a rate of precession given, for small disturbances, by the expressions

$$\text{Angular momentum} = I\omega, \quad (1)$$

$$\text{Rate of precession} \equiv \Omega = T/I\omega, \quad (2)$$

where I is the inertia of the wheel, ω is the angular velocity, and T is the torque applied at 90° to the axis of spin.

Suppose that a gyro is mounted in a vehicle, as shown in Fig. 2. In this drawing, the two gimbals on which the spinning wheel is mounted are clearly visible. As the vehicle rotates about the axis labeled *input axis* transverse to the spin axis, it exerts a torque T on the *spin axis* and there is precession about the axis labeled *output axis*. A measure of the rate of precession will be proportional to the vehicle rotation, as shown by Eq. 2. A gyro used in this manner as a rate sensor is called a *rate gyro*. To keep the displacement of the wheel within narrow limits, a restoring torque motor must be used. Further, if the rate of precession is integrated over time, a value of vehicle angular displacement is determined. A rate gyro used in this manner is termed an *integrating rate gyro*. This approach

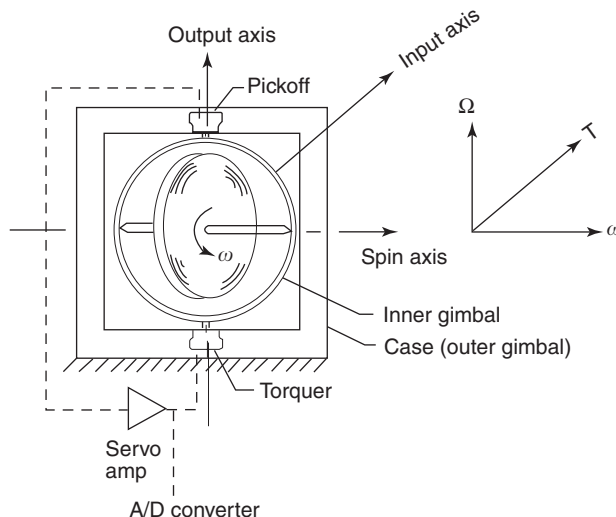


Figure 2. Gyroscopic precession.

to determining vehicular attitude is usually much more accurate than the vertical gyro approach.

When the gyro is used as a sensor, a number of design considerations come into play. First, the wheel is spun using a high-efficiency electric motor. Next, in an effort to reduce volume, the wheel may be made small and the spin rate increased to retain high angular momentum, a desirable feature for measuring very small angular rates. Next, bearings inevitably cause some disturbing torques on the wheel, and in an effort to reduce this effect, the wheel may be supported on a thin film of liquid or gas in lieu of conventional bearings. An optical pickoff is preferable to inductive pickoffs to reduce reactive forces and to measure very small rates. Finally, restoring torque commands may be in the form of pulses that provide a compatible interface with a digital computer. Some design approaches for mechanical gyros used in spacecraft can be found in Ref. 2.

Optical Gyros. Another type of sensor for measuring rotational displacement is the so-called *ring laser gyro* (RLG). This is not a gyro in the conventional sense in that there are no moving parts. The principle of operation is the Sagnac Effect discovered in 1913 (3). The sensing element is a laser beam that is split into two beams directed clockwise and counterclockwise in a somewhat circular, closed, vacuum chamber. Three or more mirrors are arranged in a “ring” around the chamber so that the two beams are reflected back to the source where there is a detector. A conceptual design approach is shown in Fig. 3. If the chamber is rotated in either direction about an axis perpendicular to the plane of the mirrors (i.e., the *input axis*), there will be a measurable difference (phase shift) in the travel times of the two beams because one will travel a shorter distance than the other. The output of the RLG can be digitized for rotational rate output and then integrated over time for a measure of angular displacement. The disadvantages of RLGs are the difficulty and cost of achieving and maintaining the necessary mechanical alignment (4).

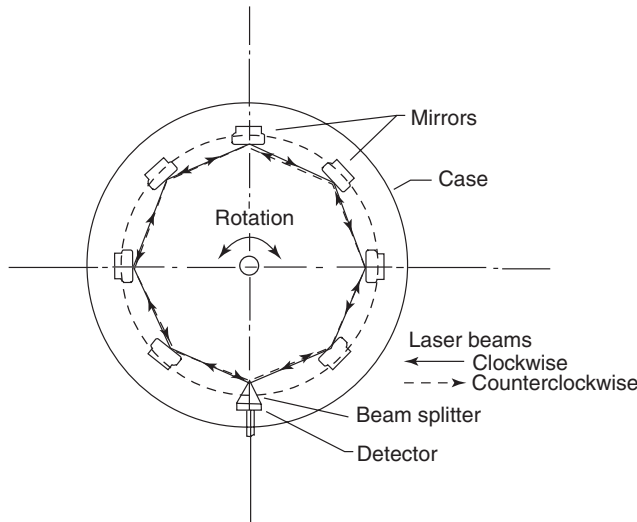


Figure 3. Ring laser gyro.

Fiber-optic gyroscopes (FOG), sometimes called *interferometric fiber-optic gyroscopes* (IFOG), operate on the same principle as the RLG in that there are two beams of light directed in opposing loops, but in this case the medium is optical fiber. A conceptual drawing of the FOG is shown in Fig. 4. In general, FOGs are easier to fabricate and align than RLGs and are more suitable for microminiaturization. One disadvantage is that scale factors in FOGs are usually nonlinear. Also, the fiber must be carefully chosen to avoid the potential of becoming unserviceable due to aging or radiation in space. FOGs have flown on the *Clementine* and *Technology for Autonomous Operational Survivability* (TAOS) spacecraft (4). An extensive discussion of the FOG may be found in (5).

Vibratory Gyros. The *vibratory gyro* is another type of gyro different from the classical mechanical gyro. The ancestry of this type can be traced to the experiments of G. G. Bryan, a British physicist, who studied vibrating wine glasses (6). He discovered that the induced vibrational pattern on a glass would move (precess) if the wine glass were rotated about its stem and that the

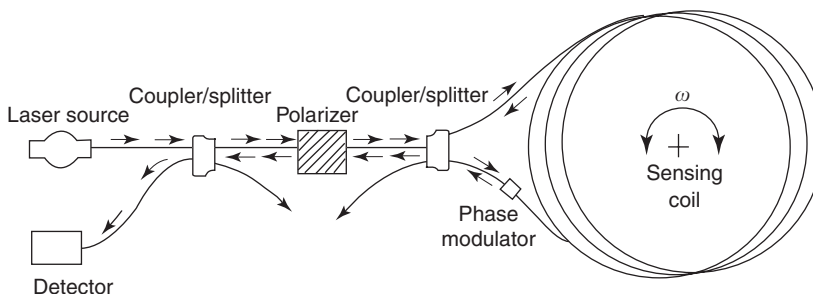


Figure 4. Fiber-optic gyro.

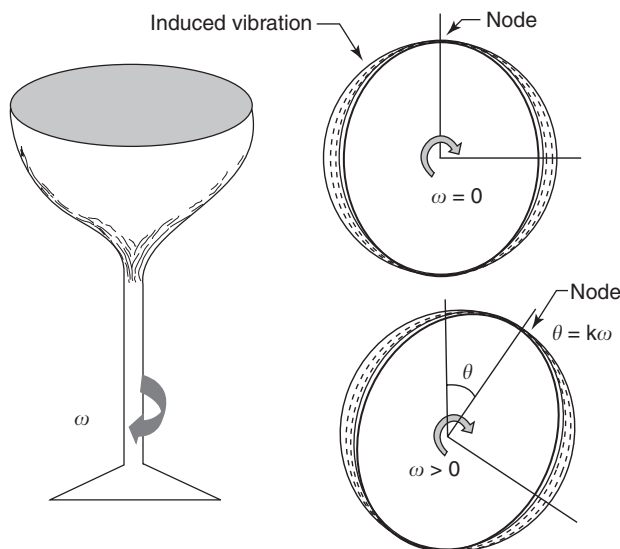


Figure 5. Vibrating bell gyro.

displacement was proportional to the rotational rate. This is shown in Fig. 5. An example of this type of gyro is the *hemispherical resonator gyroscope* (HRG) discussed in Ref. 7. In this example, the resonating element that is analogous to the wine glass is a 30-mm diameter bell made of fused silica. A surrounding housing induces vibration and also senses the nodal pattern shift through the use of capacitive pick-offs. This gyro has been used on satellites and on the Jet Propulsion Laboratory's *Cassini* spacecraft. The main advantage of the HRG is that there are no moving parts other than the resonator bell. A disadvantage is that the case must be evacuated and vacuum-sealed to prevent air damping.

The *tuning fork gyro* shown in Fig. 6 is another type of vibratory gyro. In this case, the tines are excited in the plane of the tines. As the tuning fork is rotated about an axis parallel to the tines, they tend to continue oscillating in the original plane, as shown in the vector diagram in the figure. The vector component perpendicular to the plane of the tines is proportional to the rotational rate and may be measured by capacitive or optical sensing. Materials used for the tuning fork include crystalline quartz and silicon. Crystalline quartz is a highly stable piezoelectric material suitable for micromachining (8). In the case of silicon, the fork may be part of an integrated circuit chip where the controlling and sensing electronics are designed into the chip (9).

Accelerometers. Accelerometers are devices used to sense changes in velocity. They are made in a number of ways, but the common feature of most is a mass that moves in accordance with Newton's second law. This sensing mass, sometimes called the *proof mass*, may be suspended in a number of ways and held in the neutral position by a magnetic field. As the acceleration is sensed by the mass and it begins to move, a pickoff detects the movement and sends a restoring signal through an amplifier to the restoring coil. Rather than hold the mass in a neutral position, some designs force the mass to swing back and forth

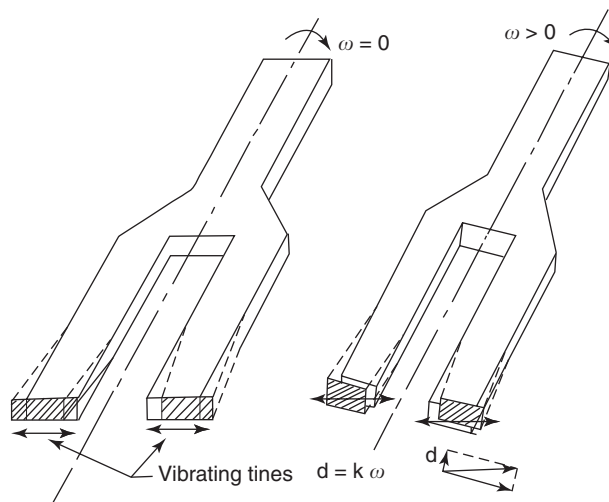


Figure 6. Tuning fork gyro.

on a pendulum using a series of back and forth pulses. This restoring circuit also sends the restoring pulses to a counter that adds the positive and negative pulses algebraically; the sum represents the sensed acceleration. If the counter is coupled with a digital computer and integrated over time, it can keep an ongoing status of vehicle velocity. This type of accelerometer, called a *pulsed integrating pendulous accelerometer* (PIPA), was used successfully in the NASA Apollo Lunar Landing Program (10).

Advancements in the past decade in *microelectromechanical systems* (MEMS), also known as *microsystems technology* (MST), have produced accelerometers of continually decreasing size, mass, and power usage (11). Using the same principles of vibratory gyros discussed before, sensing elements now include flexing quartz seismic beams and “squeezed film.” Movement may be detected by measuring changes in capacitance between the flexing mass and the adjacent fixture. *Nanotechnology*, generally defined as the next order of size reduction, will no doubt reduce the size of accelerometers further.

Inertial Measurement Units. The *stable platform*, variously referred to as the *inertial platform*, *guidance platform*, or *inertial measurement unit* (IMU) is a common application of gyros and accelerometers. In a typical approach, a group of three single-axis gyros or two dual-axis gyros is mounted on a rigid platform, and their input axes are aligned orthogonally. Three single-axis accelerometers, or two dual-axis accelerometers, are also mounted orthogonally on the platform. The platform is then mounted on two or three gimbals, and the restoring torque signals from the gyros are used to command the gimbal drive motors. The result is that, after initial erection and alignment, the platform is maintained inertially fixed in space. A platform designed in this manner provides an inertial attitude reference and measures accelerations along the inertially fixed axes of the platform. This information can be used by a flight computer to calculate and maintain the status of attitude, acceleration, velocity, and position of the platform.

The GN&C system then essentially flies the platform through space, and the vehicle moves around the platform in the process.

The three-gyro, three-accelerometer platform was used on the NASA Apollo Lunar Landing Program (10). The two-gyro, two-accelerometer IMU is currently used on the NASA Space Shuttle (12). The Space Shuttle IMU is shown in Fig. 7. For a further discussion of IMUs, see Ref. 13.

Strap-Down IMU. The limitations of early onboard flight digital computers encouraged a marriage with gimballed IMUs because of their ability to erect themselves and maintain alignment for at least short times using self-contained torquer motors and dedicated electronics. Even if the flight computer failed, the IMU alignment data could be used in these short periods of time to drive cockpit displays for manual steering. Later on, dedicated local processors within IMUs off-loaded the flight computer even more by applying scale factors, correcting biases, and encoding data words. Over the years, gimballed IMUs have earned a reputation for reliability and have become cheaper to produce.

As onboard digital computers have grown in capability there has been a trend to replace the gimballed IMU by an assembly of gyros and accelerometers rigidly mounted on the spacecraft structure. This approach is called the *strap-down IMU or strap-down guidance system*. In this approach, the flight computer must do all of the angle resolutions (body angles to inertial angles, or body angles to Euler angles) and continually maintain the inertial reference. Further, corrections must be made by the flight computer (or a dedicated local processor) to eliminate the effects of spacecraft rotation on the accelerometers. Strap-down systems are generally smaller than gimballed IMUs, require less electrical power, and are cheaper to build. A disadvantage is that they must be continually serviced by the flight computer, and if either the strap-down system or the flight computer should fail, inertial reference is instantly lost.

It is interesting to note that the Apollo Program spacecraft used a gimballed platform for the primary system and a limited form of a strap-down IMU for backup. In the latter case, body-mounted gyros were used for backup angular

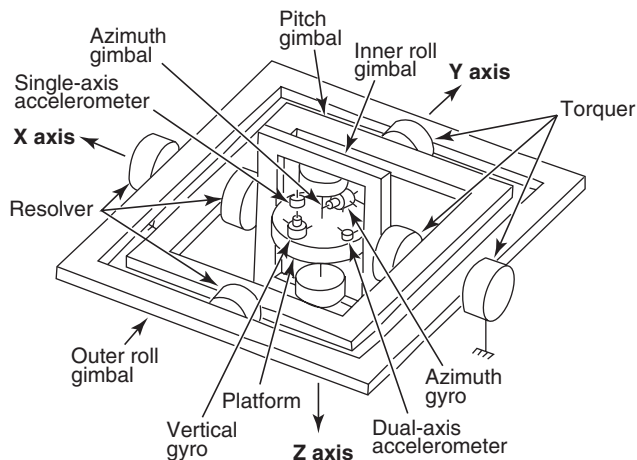


Figure 7. Space Shuttle inertial measurement unit (courtesy NASA).

rate and displacement information for both the flight computer for automatic steering and displays for manual steering. But there were no body-mounted accelerometers, and if the gimballed IMU accelerometers failed, thrust duration had to be manually timed.

It seems likely that future trends in MEMS and nanotechnology will continue to reduce the size and power usage of strap-down IMUs, making them more and more attractive for spacecraft use. However, inherent time-dependent inaccuracies in both types of IMUs require realignment using noninertial sensor or manual updating.

Rendezvous and Docking Sensors. In the hypothetical mission described before, there is a phase when the spacecraft must approach and dock with another spacecraft. Assuming that the target is passive and if the spacecraft is not manned so that manual control may be employed, the maneuver must be accomplished automatically by the onboard GN&C system. Candidate sensors for the rendezvous phase would include a Doppler radar and possibly the Global Positioning System for rendezvous in Earth orbit (14,15). A laser might be suitable for docking. Both translational and rotational commands to the spacecraft attitude control rockets are required and would be generated by the guidance function of the GN&C system. The guidance algorithm must be carefully scripted so that the spacecraft is slowed down enough to prevent impact damage but still has enough kinetic energy on impact to overcome the resistance of the latching mechanism.

Space Navigation

The foundation of space navigation was laid in the seventeenth century by two major advances. Early in the century, Johann Kepler, using the observations of Tycho Brahe, empirically derived his laws of planetary motion. The first law, and the most important for celestial navigation, stated that the planets of the solar system move about the Sun in elliptical orbits and the Sun is at one focus of the ellipses. Later, Sir Isaac Newton stated his laws of motion and formulated the law of universal gravitation. His work confirmed Kepler's findings and allowed extension to celestial bodies other than planets, for example, comets, and to motions described by conics other than ellipses. These trajectories include circles (a special case of the ellipse), parabolas, and hyperbolas. A discussion of the historical background of this development can be found in Ref. 16. (See also article Earth Orbiting Satellite Theory by S. Nerem in this *Encyclopedia*.)

Newton's law of universal gravitation may be stated generally mathematically as

$$F = \frac{Gm_1m_2}{s^2}, \quad (3)$$

where F is the magnitude of the force of attraction, m_1 and m_2 are the masses of the two bodies, s is the distance between them, and G is the gravitational constant whose numerical value depends on the system of units used. The force F points in the same direction as the line s that joins the two masses.

Equation 3 may be applied with reasonable accuracy to a spacecraft orbiting Earth if certain simplifying assumptions are made:

1. Let m_1 represent the mass of Earth; Earth is of uniform density and is spherically symmetrical, that is, the oblateness of Earth is ignored. This allows the Earth to be treated as a point mass at its center.
2. Let m_2 represent the mass of the spacecraft, so small relative to the mass of Earth that the center of mass of the system lies at the center of Earth.
3. Let m_1 be fixed in inertial space with the origin of the reference axes at its center.
4. The spacecraft is in coasting flight with only gravity acting on it, that is, other forces such as aerodynamic drag, solar winds, and electromagnetic forces are ignored.

These assumptions allow simplifying the analysis to what is generally termed the “restricted two-body problem,” and the approach may be used for spacecraft operating near other relatively large bodies, for example, where m_1 represents the Sun, Moon, or one of the planets.

Following the approach of Mueller in Ref. 17, the spacecraft equations of motion can be derived as follows. Using polar coordinates and vector notation, we show the Earth–spacecraft coordinate system in Fig. 8. Referring to Newton’s second law and restating Eq. 3 in vector form in the polar coordinate system,

$$\sum \vec{F} = m_2 \frac{d^2 \vec{r}}{dt^2}, \quad (4)$$

$$\sum \vec{F} = (-) \frac{Gm_1 m_2}{r^2} \hat{r} = (-) \frac{Gm_1 m_2}{r^3} \vec{r}, \quad (5)$$

where r is the distance between the masses, \hat{r} is a unit vector that points along the line joining the two masses.

Combining Eqs. 4 and 5 and simplifying produces

$$\frac{d^2 \vec{r}}{dt^2} + \frac{\mu \vec{r}}{r^3} = 0, \quad (6)$$

where μ is defined as a constant $\mu = Gm_1$. Equation 6 is the vector differential equation of motion for the restricted two-body problem. Note that it is independent of m_2 .

Using the methods described elsewhere in this *Encyclopedia* (see article Earth Orbiting Satellite Theory by S. Nerem) and referring to Fig. 8, two equations of importance for an orbiting spacecraft can be derived from Eq. 6:

$$E = \frac{v^2}{2} - \frac{\mu}{r}, \quad (7)$$

and

$$H = rv \cos \gamma, \quad (8)$$

where γ is defined as the *flight path angle*.

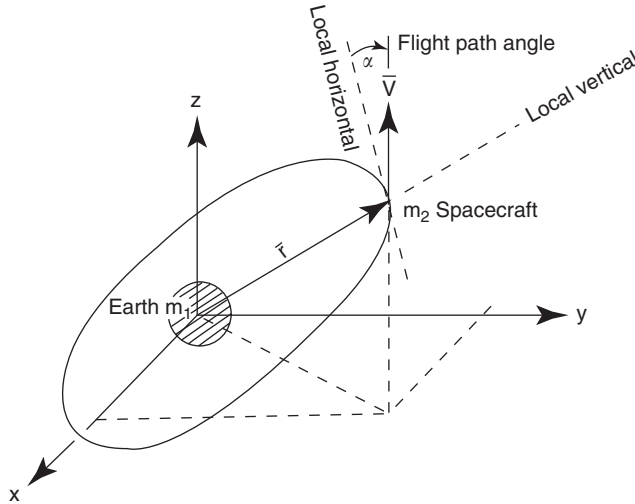


Figure 8. Earth-spacecraft coordinate system.

E is a constant in Eq. 7, termed the *specific mechanical energy* of the spacecraft, implying a continual exchange of kinetic energy and potential energy throughout the orbit. H in Eq. 8 is called the *specific angular momentum* and is also constant throughout the orbit. Notice that as r increases, v decreases, as might be expected intuitively.

Finally, referring again to the article by S. Nerem, Eq. 6 can be solved, resulting in the following scalar equation termed *the trajectory equation*:

$$r = \frac{H^2/\mu}{1 + B/\mu \cos v}, \quad (9)$$

where, as before, H is the specific angular momentum and B is the magnitude of a vector from the focus to that point on the trajectory nearest the focus. This point is called the *perifocus* or *periapsis*, or in the case of Earth orbits, the *perigee*. The angle v (nu) is the angle between \vec{B} and \vec{r} , the radius vector.

Equation 9 is in the form of the general equation for a conic section written in polar coordinates. If p is defined as $p = H^2/\mu$ and e is defined as $e = B/\mu$, the equation can be written as

$$r = \frac{p}{1 + e \cos v}. \quad (10)$$

In this form, p is called the “parameter” or “semilatus rectum” and is a measure of the size of the conic. The term e is called the “eccentricity.” The solution of Eq. 10 is a circle if $e = 0$, an ellipse if $0 > e < 1$, a parabola if $e = 1$, and a hyperbola if $e > 1$.

In summary, application of these laws to a spacecraft operating in a central force field results in trajectories that are similar to those of celestial bodies. A coasting spacecraft that has sufficient energy will orbit Earth in a plane and in

an elliptical fashion. With the additional thrusting it can be made to travel further out to loop around the Moon and return to Earth, or be captured in an orbit about the Moon. Even more thrusting will cause it to escape Earth's gravitational pull and proceed on a parabolic or hyperbolic path to one of the other planets or into deep space. The resulting trajectories can be divided into conical segments, whose combinations are called *patched conics*.

As mentioned earlier, limitations in onboard expendables require planning most space trajectories carefully in advance of the actual flight to achieve the most economical and efficient missions. Launch dates, orbital plane changes, midcourse velocity changes, and rendezvous points must be determined by working backward from a desired target along the way and at the end of the flight. This is usually done by using numerical integration with considerable trial and error adjustments.

Once the mission plan is determined, navigational sightings are defined in terms of times, locations, and types of sensors to be used. At each point, the sightings are made, the information is routed to the onboard flight computer, and the inertial platform is aligned. If an adjustment is required in the state vector of the vehicle, attitude alignments of the vehicle are made and midcourse delta V 's are made using the thrusters on the vehicle. This is repeated as often as required to achieve mission goals.

On the launch pad, the inertial measurement unit is held in a locked position relative to the spacecraft until the last practical moment, usually a few minutes before ignition. In the minutes after IMU release to actual vehicle liftoff, the IMU is controlled by a gyrocompass program to keep it aligned relative to Earth. At liftoff, the IMU is allowed to go inertial and remain so until the next navigational update in flight.

During flight, there are several types of navigational updates. Ground radar tracking is the most common for spacecraft orbiting Earth. Optical sensors may be used to take sightings of Earth, the Sun, or the stars. These sensors include star trackers, horizon scanners, or Sun seekers for automatic navigation. For the manned NASA Apollo Lunar Landing Program, the astronauts made visual sightings using a telescope and sextant that were coupled electronically to the flight computer. All of these sensors are usually carefully mounted on a rigid navigational base that also supports the inertial measurement unit, so that angular resolution from an optical sensor to stable member axes can be made precisely.

Navigational fixes can also be made using the *Global Positioning System* (GPS) satellites [see also the article on Global Positioning System (GPS) elsewhere in this *Encyclopedia*]. This approach can be used to determine the vehicle attitude as well as the location in space if multiple receivers are located precisely on the spacecraft and their relative positions are differentiated (15).

Star Trackers. The *star tracker* is an automatic optical device used to determine the angle between the spacecraft and a luminous body typically outside the solar system. Planets do not make good targets because they lie in a fairly narrow band (the zodiac) and their motion is erratic compared to stars many light-years away. A candidate group of stars is usually preselected and stored in the flight computer along with their general location and their brightness number. The spacecraft is oriented so that the star tracker points in the general direction

of a candidate star and then searches until a match is made. Alternately, the tracker itself may be gimbaled and allowed to move relative to the spacecraft axes. Two different sightings are enough to establish the spacecraft's position in space or to align the IMU stable member, but a third or more additional fixes are used for confirmation and greater accuracy. Star trackers now in production have angular accuracy of 0.1 arc seconds or less (18). Star trackers are sometimes used to track other spacecraft, and they may be combined with cameras for photography.

Horizon Scanners. The *horizon scanner* is used to determine the *local vertical* of a spacecraft that is orbiting Earth or other planetary bodies. The local vertical may be considered a vector from the center of mass of the vehicle to the center of Earth. Three or more sightings are taken of Earth's horizon, as shown in Fig. 9, and the angles to the local vertical relative to the spacecraft body axes or inertial axes are computed geometrically by the flight computer. Once the local vertical is determined, star sightings can be made and the latitude and longitude of the vehicle determined. Since visible light is scattered by Earth's atmosphere, the sensor is usually designed to detect infrared waves that more sharply define Earth's horizon. Another advantage of infrared is that Earth radiates heat at infrared wavelengths even when the horizon is not in the direct rays of the Sun, and, consequently, sightings can be made on the dark side of Earth.

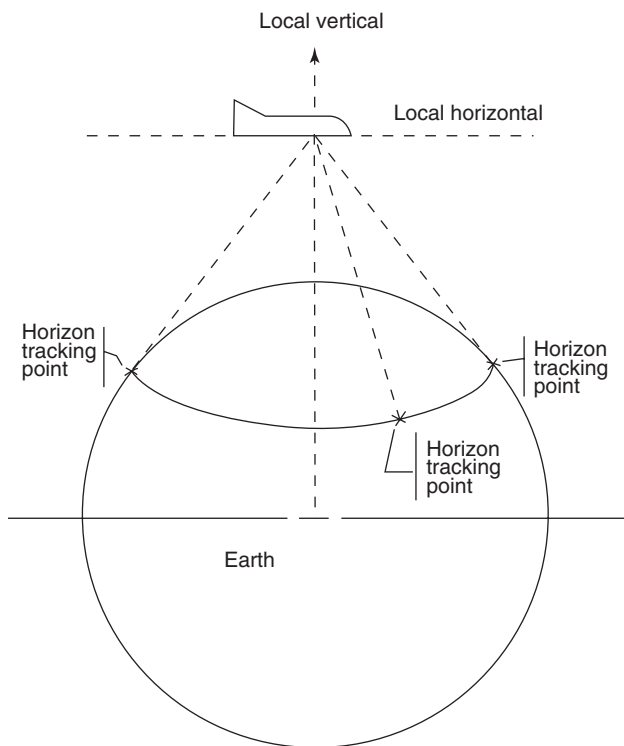


Figure 9. Horizon scanner.

Sun–Earth Sensors. *Sun–Earth sensors*, sometimes individually referred to as *Sun seekers* or *Earth sensors*, are sometimes used as attitude determination devices for spacecraft. They are relatively simple to build, and sightings are usually reliable because the Sun and Earth are large targets and hard to miss. But for the same reasons and because of the relatively rapid movement of Earth in its orbit, they are not usually as accurate for navigation as other devices previously discussed. A principal use of Sun seekers is to orient the vehicle relative to the Sun for thermal control.

Radio Navigational Aids. For Spacecraft required to return safely to Earth and land (e.g., the NASA Space Shuttle Orbiter) or land on other bodies, one or more radio navigational aids may be required. The simplest of these is the *radar altimeter*, which allows the spacecraft to measure altitude autonomously above the surface of the landing zone. In the Space Shuttle Orbiter, the radar altimeter measures altitude from about 5000 ft down to touchdown.

Tactical Air Navigation (TACAN) units are used on the Space Shuttle Orbiter to determine the slant range and magnetic bearing of the Orbiter during landing approach. This is not an autonomous capability, and several precisely located active ground stations are required. The maximum range of TACAN is about 400 nautical miles. The Orbiter acquires bearings at a range of about 300 nautical miles and 160,000 ft altitude after entry blackout. TACAN data can be used down to an altitude of about 1500 ft.

For final approach, the Orbiter uses the *Microwave Scanning Beam Landing System* (MSBLS). This system requires active ground stations located immediately adjacent to the runway. The Orbiter acquires the MSBLS signal at a range of 8 to 12 nautical miles and at an altitude of approximately 18,000 feet. The Orbiter Commander usually takes over manually over the runway threshold at about 100 feet altitude.

Information on the NASA Space Shuttle is taken from Ref. 12.

Control

A number of functions must be performed during spacecraft flight that fall under the general heading of *control*:

- During navigational observations, the spacecraft must be aligned relative to an inertially fixed axis system and the attitude stabilized within a very narrow angular deadband. This is sometimes called the *attitude hold mode*.
- During periods of thrusting, the spacecraft must be pointed in the correct direction, and the thrust vector controlled. This is usually called the *delta V mode*.
- Spacecraft attitude hold relative to the Sun may be required, or perhaps the vehicle is slowly rolled (so-called “barbeque” mode) for thermal control.
- Attitude control may be required in conjunction with the deployment of certain mechanisms such as solar panels, radiators, docking mechanisms, and manipulator arms.
- If the spacecraft is to land on Earth or on one of the planets that has an atmosphere, control during atmospheric entry may be required and may

necessitate blending of attitude control rockets and aerosurfaces. Landing control is likely to be required and may include control of aerosurfaces, speed brakes, parasails, drag parachutes, landing gear, wheel brakes, and steering on the ground.

- Effector command signals must be processed (scaled, mixed, prioritized, time-delayed) and routed.

Depending on the GN&C system design, attitude change commands may come from the flight computer, directly from the flight crew, or from the ground via radio links. Attitude stabilization commands may come from body-mounted gyros that are generally less accurate than those used for inertial guidance but may serve as emergency backups for the IMU. A newer technique for stabilizing spacecraft in Earth orbit uses differential Global Positioning System (GPS)-derived position data from multiple receivers located remotely from each other on the spacecraft. It has been found that attitude knowledge of the order of 0.05° is possible (15). Body-mounted accelerometers may be used for docking sensing or aerodynamic drag sensing during entry.

Effectors. Devices that produce intentional reactive forces on the spacecraft are termed *effectors* and may include any of the following:

- Attitude control thrusters that use cold gases or reactive chemicals as propellants. These may also be used for small translations for such maneuvers as docking.
- Major engines that produce large changes in the velocity of the spacecraft (ΔV). Both thrust level and direction (*thrust vector control*) may be controlled.
- *Reaction wheels* where the wheel rotational rate is accelerated or decelerated to achieve reactive torquing of the spacecraft and a corresponding change in attitude.
- *Control moment gyros* that are typically mechanical gyros torqued electrically so that the resulting precession produces a desired change in spacecraft attitude.
- Aerodynamic surfaces and drag devices.
- Tethers that produce electromagnetic thrusting.

Usually the effectors themselves are not considered part of the GN&C system, but their control electronics are. Analysis of the effects on vehicle motion is usually considered a GN&C responsibility. A simplified generalized block diagram of control function is shown in Fig. 10.

Environmental Disturbances. There are several disturbances that can cause a variation in attitude and possibly tumbling or wobbling. These are usually more noticeable during quiescent periods when the spacecraft is allowed to drift:

- Gravity gradient effects, usually significant when orbiting or flying near a large body.

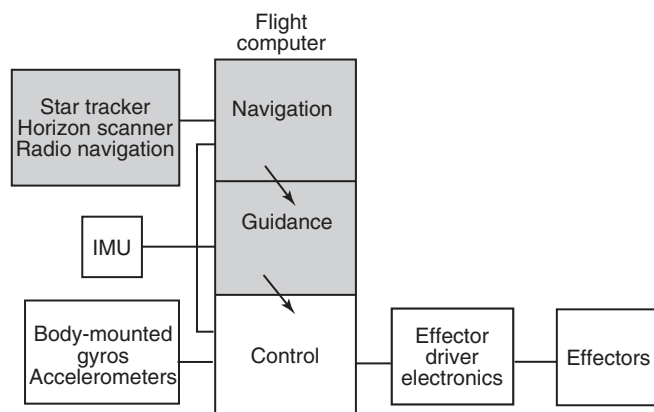


Figure 10. Control block diagram.

- Aerodynamic drag and moments when flying near a body having an atmosphere.
- Solar radiation pressure, usually significant when the spacecraft configuration includes large planar surfaces, such as solar cell panels.
- Electromagnetic induction when flying through the magnetic field of a large body.

The effects of these disturbances are difficult to calculate accurately but may be estimated using the methods of Refs. 19 and 20.¹ At the appropriate time, when a fixed attitude is required, these must be counteracted by the attitude control system. Often a small angular deadband is desired and is maintained by use of the reaction wheels, control moment gyros, or attitude control thrusters. In the last case, thruster propellant usage is always a consideration and sometimes leads to sophisticated electronic logic for duty cycle optimization.

Reference Axis System. Since effectors and some sensors are mounted rigidly to the spacecraft structure, it is customary for control analysis to use an axis system originating at the center of mass of the spacecraft. This is depicted in Fig. 11. Symbols are defined in Table 1, followed by a sign convention in Table 2.

Equations of Motion. Development of the equations of motion for all mission phases of a spacecraft is an arduous task and beyond the scope of this article. The usual approach to analyzing a particular phase is to make certain simplifying assumptions to reduce the number of mathematical terms significant in that phase. Later on, after achieving a basic understanding of the dynamics, correcting terms may be added for evaluation.

¹Another approach might be that used by the U.S. Navy to estimate aerodynamic stability derivatives in the transonic region for aircraft. Displaying remarkable initiative and refreshing candor, the Bureau of Aeronautics claims success using the method of omphaloskepsis (see BuAer Report AE-61-4, Fundamentals of Design of Piloted Aircraft Flight Control Systems, Volume II, Dynamics of the Airframe, February 1953, p. v-11.)

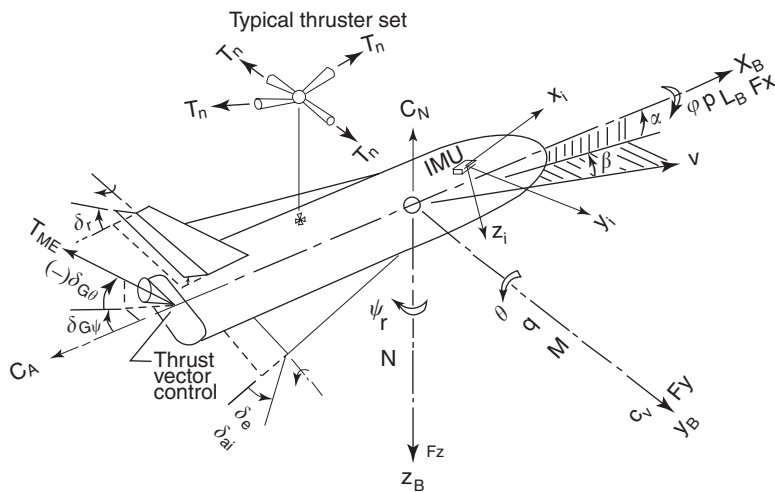


Figure 11. Control reference system.

For example, shown below are *six-degree-of-freedom* equations of motion for the coasting (in space) attitude hold mode. Simplifying assumptions are as follows:

- The spacecraft is considered a rigid body.
- Vehicle mass and inertia remain constant during the period of interest.

Table 1. **Body Axes Coordinate System**

x_B, y_B, z_B	Orthogonal body axes, right-hand rule
x_i, y_i, z_i	Orthogonal inertial axes, right-hand rule
F_x, F_y, F_z	Forces along the body reference axes
L, M, N	Moments about the x_B, y_B, z_B axes, respectively, right-hand rule
φ, θ, ψ	Angular displacement about the x, y, z axes, respectively
p, q, r	Angular velocity about the x_B, y_B, z_B axes, respectively
V	Velocity vector
α	Angle between the x axis and the projection of the velocity vector in the x - y plane; angle of attack
β	Angle between the x axis and the projection of the velocity vector in the x - z plane; angle of sideslip
$\delta_{G\theta}$	Main engine gimbal deflection (pitch)
$\delta_{G\psi}$	Main engine gimbal deflection (yaw)
δ_e	Elevon deflection
δ_a	Aileron deflection (produced by differential elevon deflection)
δ_r	Rudder deflection
$T_n \ n = 1, 2, 3 \dots$	Attitude control thrusters; each has a unique number and may produce forces along all body axes
T_{ME}	Main engine thrust; may produce forces along all axes and moments about all axes

Table 2. Sign Convention

Cause	Effect
(+) angle of attack (+ α)	(-) z airload
(+) sideslip (+ β)	(+) y airload
(+) rudder deflection (+ δ_r)	(+) y force, (-) yaw (- N)
(+) elevon deflection (+ δ_e)	(-) z force, (-) pitch (- M)
(+) differential elevon deflection (+ δ_a)	(-) roll (- L)
(+) gimbal deflection in pitch (+ $\delta_{G\theta}$)	(-) z force, (-) pitch (- M)
(+) gimbal deflection in yaw (+ $\delta_{G\psi}$)	(+) y force, (-) yaw (- N)
(+) force	(+) linear acceleration
(+) moment	(+) angular acceleration
(+) attitude control thruster force (+ T_n)	(+ or -) moment depending on thruster location
(+) C_N normal force coefficient	(-) force along the z_B axis
(+) C_A axial force coefficient	(-) force along the x_B axis
(+) C_Y side force coefficient	(+) force along the y_B axis

- The spacecraft is symmetrical about the x - z plane, causing the products of inertia J_{yz} and J_{xy} to drop out.
- Terms involving the products of inertia, pJ_{xz} and rJ_{xz} , are small and may be ignored.

$$\ddot{x} = 1/m \left[\sum F_{x \text{ THRUSTERS}} + \sum F_{x \text{ ENVIRONMENT}} \right],$$

$$\ddot{y} = 1/m \left[\sum F_{y \text{ THRUSTERS}} + \sum F_{y \text{ ENVIRONMENT}} \right],$$

$$\ddot{z} = 1/m \left[\sum F_{z \text{ THRUSTERS}} + \sum F_{z \text{ ENVIRONMENT}} \right],$$

$$\ddot{\phi} = 1/I_{xx} \left[\sum L_{\text{THRUSTERS}} + \sum L_{\text{ENVIRONMENT}} \right],$$

$$\ddot{\theta} = 1/I_{yy} \left[\sum M_{\text{THRUSTERS}} + \sum M_{\text{ENVIRONMENT}} \right],$$

$$\ddot{\psi} = 1/I_{zz} \left[\sum N_{\text{THRUSTERS}} + \sum N_{\text{ENVIRONMENT}} \right].$$

It is helpful to write the equations in this form to get an understanding of the importance of the different terms. Numerical solution on digital computers is usually more convenient after conversion to matrix form. For further discussion on this subject, see Refs. 13, 21–23.

The Flight Computer

There is no part of GN&C technology that has evolved during the past 50 years as dramatically as the onboard digital flight computer. It started in the early

1950s as a fire control computer 4 cubic feet in volume with 250 vacuum tubes (24). Then came the addition of the onboard navigation function with transistor-transistor logic (TTL). Then its use was expanded to include guidance where steering signals were injected into an analog flight control system. By the time of the second-generation Apollo spacecraft, the flight control function was added, integrated electronic circuits were implemented, and the first automatic fly-by-wire flight control was realized. With the Space Shuttle came computer control of all GN&C functions for both atmospheric and spaceflight, computer-driven cockpit displays and computer processing of all manual flight control inputs, and the beginnings of distributed processing. Current flight computers are on the order of cubic inches in volume or smaller, use relatively modest amounts of electrical power, and compute at rates incomprehensible only a few years ago. Microminaturization is being realized in current applications, and the application of nanotechnology seems likely within a few years.

Today's flight computer performs an impressive array of functions, depending on the mission of the spacecraft:

- Guidance sensor management and data processing, guidance algorithm computation, and steering signal generation.
- Navigation sensor management and data processing; course evaluation, correction, and forecasting; and trajectory calculation.
- Flight control sensor management and data processing, vehicle attitude control, command mixing and prioritization, effector selection, thrust vector control, attitude thruster control, aerosurface control, vehicle bending and longitudinal oscillation control.
- Aerobraking control, parachute and parasail deployment and steering, nose wheel steering, and wheel braking.
- Systems management for non-GN&C systems, such as electrical power generation, distribution and control; radio/radar/television communications; exploratory payload sensor management and data processing; environmental control systems; and hydraulic and pneumatic systems.
- Consumables accounting and management.
- Flight instrumentation control, data processing, and downlink control.
- Generation of cockpit displays and processing of manual commands for flight control.
- Accommodation of ground commands via radio for software updates, GN&C commands, and data downlinking.
- Redundancy management.
- Vehicle health management (an extension of onboard checkout).
- Multiplex data bus management.
- Distributed processing management.

Input/Output. Probably the most difficult problem in implementing flight computers is communication between the computer and the various devices that are commanded or generate data. The magnitude of this problem can be appreciated intuitively if one imagines the number of wire bundles and

connectors required to link the devices to the computer. The problem is compounded if there are redundant computers linked to one another that share the same information. Also, put simply, the digital flight computer is an anomaly in the middle of an analog world, and the necessary conversion of signals from analog-to-digital (A/D) and digital-to-analog (D/A), plus attendant voltage scaling and device scale factor and bias accommodation, is a significant task.

The computer communications problem has been solved successfully in various ways:

- Separating the input/output function and the computation function into two boxes called the input-output processor (IOP) and the central processor unit (CPU). In this approach, the IOP handles the A/D and D/A conversions, voltage scaling, temporary data storage, data bus management (if required), and other functions.
- Doing the A/D and D/A conversion, scale factor adjustment, and voltage scaling at the devices served. In some cases, small local processors are implemented in the devices to perform these relatively simple chores, so that messages to and from the main flight computer are reduced to significant flight data. (This is the beginning of *distributed processing*.)
- Implementing a multiplex data bus distribution system for communicating with the various devices. In this approach, various devices that have unique addresses are connected to a data bus managed by the IOP. An important advantage of this approach is weight reduction in wiring.
- Sharing the computation task among several processors on the spacecraft. It could be argued that this *distributed processing* approach is more of a fundamental change in the approach to hardware and software than just a solution to the I/O problem. Certainly, it significantly affects the total software design, implementation, and verification/validation approach. In addition, it usually reflects not just a GN&C decision, but total spacecraft system engineering methodology.

Real-Time Operation. The requirement of uppermost importance for flight computers doing GN&C computations is the ability to keep up with the dynamics of the vehicle and the sensors that provide input data. Whereas batch processing computers in typical ground settings simply run longer when necessary to complete a job, such a delay can be deadly in flight computers in the loop of highly dynamic flight scenarios. The flight computer must maintain system stability, provide computational precision, service sensors, accept interrupts, cope with failures, and even monitor itself while running on control software cycle times typically of 40 milliseconds. The computation bandwidth is almost impossible to estimate accurately early in the design phase of a program, and even the most generous growth margins are usually exceeded and call for compromises to be accommodated later on in the flight software.

A key characteristic of the computer loading is the iterative nature of many calculations, particularly in navigation. There will be errors from a number of sources, including navigation sensor measurements and system mechanization.

A technique for handling these errors is the Kalman filter (25,26), a stochastic analysis and estimation process used extensively in space, aircraft, and missile computers since the early 1960s. Beginning with a priori knowledge of system errors, the Kalman filter continually performs a statistical error analysis and predicts new values for system variables. The resulting product is continuing improvement in position determination.

Flight Software. In the early days of general-purpose digital computers, and especially in digital flight computers, computation times were extremely slow by today's standards, and random access memory (RAM) was bulky and expensive. The typical computer programmer was a mathematician who likely had a keen appreciation of the overall computation objectives and was driven to code austere programs in what was termed *machine language* in those days but called *assembly language* today. Major disadvantages in that software coding approach were that it was labor-intensive and only the original programmer understood the logic behind his organization of the computer program. If other programmers were brought in to modify the program, there was usually some unproductive period to learn the code already in existence. The major advantage was that the assembly code was understandable, at least to the original programmer, and was easily modified. The GN&C engineer often simply had only to speak in general terms to the programmer to have a change incorporated in a timely manner.

Since those early days, three things have happened to cause a revolution in the way flight software is produced today: computation rates have increased by orders of magnitude, random access memory has become inexpensive, and *higher order languages* (HOL) have become predominant in the development of software programs. Now, software engineers and programmers are formally educated in information systems technology rather than mathematics. The use of higher order languages such as FORTRAN, C, C++, Ada, and HAL/S makes the flight software engineer and programmer very productive in terms of assembly code generated. And once software engineers or programmers learn the rules and methods for a particular HOL, they can become productive immediately without knowing very much about the "big picture" of the flight program.

This evolution in computer hardware and software production is not without drawbacks. The very universality of a typical HOL that makes it attractive to modern diverse users also produces very inefficient assembly code. Traditionalists cringe at the squandering of RAM, and even today there never seems to be enough RAM as a spacecraft program progresses. Then, there is the assembly code itself—often virtually indecipherable and difficult to change if relatively simple modifications (called *patches*) are necessary. The alternative is to make the changes in the source code in the HOL and recompile, a time-consuming process and so prone to introducing undesirable effects that lengthy reverification is usually required. An unfortunate result of this modern approach is that the GN&C engineer usually does not have a thorough knowledge of the flight software program and has difficulty developing an intuitive "feel" for the program. He must rely almost exclusively on the formal verification/validation of the program for confidence.

Another characteristic of the modern GN&C flight software program is that it seems to "grow like Topsy" as the spacecraft design matures. This is an

unfortunate product of an evangelistic-like effort in the early days of spaceflight, and still today, by enthusiasts to “sell” the digital flight computer by welcoming more and more functions into the software program as a cheap way of implementation. Software accommodation of a feature may appear initially to be a simple and seemingly inexpensive thing to do, but often the penalties in time and expense for the subsequent verification/validation are not anticipated. A vigorous GN&C system engineering program is probably the best solution for this quandary.

The cost of flight computer software is likely to be the most significant item in the GN&C budget and may exceed the cost of the rest of the GN&C system. There are several important factors involved in containing this cost:

- It is critical to document software requirements in as much detail as possible or the computer may be undersized, or the software cost driven to astronomical levels by changes later on, or both.
- The HOL must be chosen carefully while balancing its suitability, maturity, and compilation time with the availability of software engineers and programmers familiar with it.
- The executive program (operating system) must be carefully planned for real-time operation with the attendant issues discussed earlier.
- Application software must be compatible with the devices commanded, and the software engineer-to-device engineer interaction is expensive and sometimes difficult to orchestrate.
- All software must be designed recognizing the time, manpower, and equipment costs associated with the verification process.
- Configuration management is an absolute must for the duration of the spacecraft project.
- The verification and validation approach must be selected. For verification, will flight computers or computer emulators be used? For validation, will a flight computer plus simple simulator be used, or will an “iron bird” or high-fidelity avionics integration facility plus high-fidelity simulator be used? And will performance of the verification/validation be done (or repeated) by an independent organization (so-called *independent verification and validation* or IV&V)?

Current Trends. Flight computers continue to grow in capability—faster computational speeds and greater RAM—and are becoming smaller, less power hungry, and cheaper. Choices that were significant a few years ago are not issues anymore. For example, fixed-point versus floating-point arithmetic is now typically decided in favor of floating-point even though floating-point takes more computing time. Word length is now typically 32 bits, and 16-bit machines are passing from the scene. Lasers are being used for communication within the computer in place of copper wire. More and more devices in the GN&C system have embedded digital processors that take care of much of the computer overhead, thus off-loading the central flight computer. Other advancements are discussed under “Integrated GN&C” following.

Integrated GN&C

Modern spacecraft GN&C systems have their roots in the automatic pilots (*autopilots*) developed for aircraft in the middle of the last century to relieve pilots on long flights. Rapid advances were made in the 1930s and 1940s, inspired largely by World War II. Initial, relatively crude, vacuum-driven gyros plus vacuum control valves plus hydraulic actuators gave way to electrical sensors plus electronic (vacuum tube) signal processing plus electrical actuators. Throughout this period, the guidance and navigation functions were typically performed by the flight crew and manual commands could be issued to the aircraft directly or via the autopilot control system. Autopilots were the means used to control bombers with steering commands coming from the famous Norden bombsight. Conceptually, then, the guidance and navigation functions were an “overlay” to the autopilot control system.

During the 1950's to early 1960's when onboard digital flight computers came on the scene to perform the G&N functions this overlay approach was followed. Flight computers provided steering commands to the vehicle via the analog control system in much the way aircraft crews had done in the past. The analog control system continued to provide three-axis stability and processed the computer-generated steering commands. This design approach was typical of aircraft and guided missiles of the period, and it was followed on the NASA Gemini spacecraft and the early NASA Block I Apollo spacecraft. It is interesting to note that toward the end of this period when manned spacecraft came on the scene mechanical control linkage gave way to electrical communications between the astronaut and effectors and the expression *fly-by-wire* originated.

In the mid-1960s, as flight computers became more powerful and compact, the entire GN&C calculation function was assigned to them, and analog channels were retained as backup. Outputs were then made directly to the driver amplifiers of the analog control system to command the attitude control thrusters and large engines for changes in translational velocity and thrust vector control during thrusting. This *integrated* GN&C System became the standard design approach used in guided missiles, satellites, the NASA Apollo lunar mission spacecraft, the NASA Space Shuttle, and is followed today in the International Space Station. This evolution is shown in Fig. 12.

The *integrated* approach has several advantages. Since the flight computer performs all the GN&C calculations, the computer program (“flight software”) development can be managed more easily. Flight software design changes can be controlled and implemented with less chance of error. System internal redundancy, if desired, can be implemented more efficiently. Overall GN&C electrical power usage is less. Total system volume and mass are reduced.

This is not to say that the integrated GN&C System is necessarily less complex. Likely, it is not. Computer programs become larger, harder for nondeveloper users to understand, and almost impossible to test completely. Training of flight crew and ground support personnel is usually more tedious and lengthy. Preflight testing of the integrated system becomes more involved and expensive because of the fidelity required.

Nonetheless, the integrated GN&C System provides the capability of performing more complex missions and is more flexible for design changes

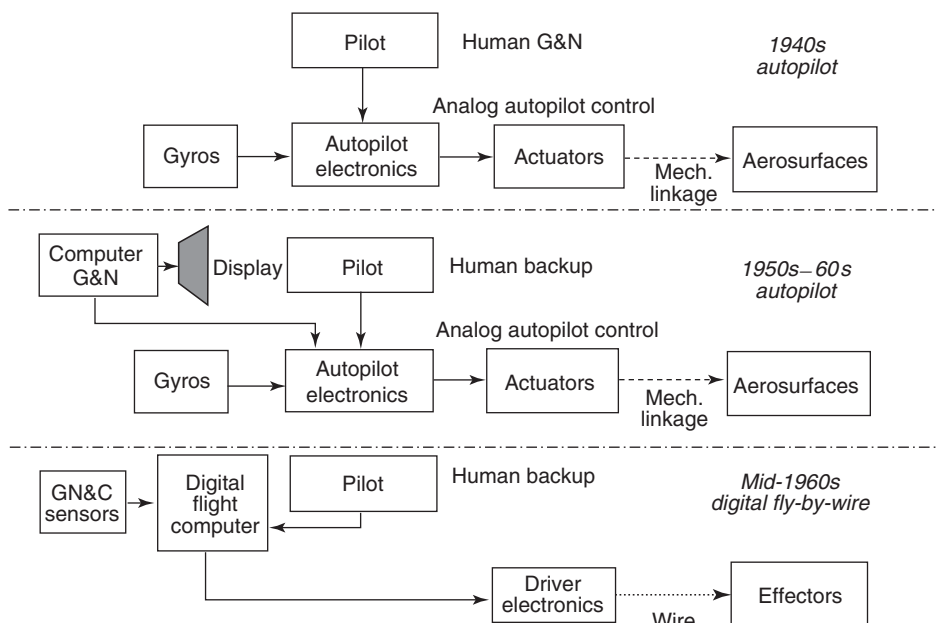


Figure 12. Evolution of the autopilot.

needed for corrections, evolutionary development, and last minute mission changes.

Digital Data Buses. A major advancement in the design of the NASA Space Shuttle was the use of multiplex serial digital data buses for communication with the flight computer. This approach helped solve the computer input/output loading problem and also provided substantial weight saving. Some 28 data buses are used in the Shuttle design. The data bus is, physically, a twisted shielded pair of wires with transformer couplers for bus protection at each electronic "user" box. The computer uses command and data words that have unique addresses for each receiver so that although all boxes on the bus hear all communications, each responds only to words that have its address. Multiplexers/Demultiplexers convert and format serial digital commands into separate parallel discrete, digital, and analog commands for the various systems served (demultiplexing) and do the reverse for data collected and sent back to the computer (multiplexing). The USAF continued this development, and multiplexed data buses are now common on aircraft, spacecraft, and missiles (27).

An approach similar to the copper data bus discussed before is the fiber-optic bus (28). Fiber has the advantages of potentially higher data rates and less susceptibility to electromagnetic interference or intentional jamming. Disadvantages include larger minimum bend radius for installation and the potential to become unserviceable due to age or environmental effects.

Fiber optics is also considered part of a general class of communications called *photonics*. This is a term used for the general case in which the medium is

light rather than electron flow. In GN&C, the expression *fly-by-light* is sometimes used. When direct line of sight exists between devices, the use of laser beams or infrared beams for data transfer is possible.

Fly-by-Wireless? A promising development that could be used in lieu of the data bus is digital spread-spectrum radio frequency communication between the flight computer and devices in the GN&C system. *Spread-spectrum* is a technique used to reduce or avoid interference by taking advantage of a statistical means to send a signal between two points using different frequencies at different times. The theory is that noise tends to occur at different frequencies at different times. Therefore, even though part of a transmission might be lost due to interference, enough of the message will come through to create noticeably better output compared to fixed-frequency systems. Further, using error correction techniques, the original message can be totally restored. The promise of this approach for aircraft and spacecraft is reduction in size, weight, and power of GN&C systems while offering immunity to natural interference or jamming from man-made equipment.

INVOCON, based in Conroe, Texas, has had considerable success in developing instrumentation for the NASA Space Shuttle Orbiter and the International Space Station by linking sensors to a central controller/transponder via *spread-spectrum radio*. Using the same principle on a project sponsored by NASA Dryden Flight Research Center, INVOCON successfully replaced a data bus link from the flight computer to an elevon actuator on a NASA F-18 research airplane (29). This concept is shown in Fig. 13.

Redundancy. The use of backup hardware in aircraft and spacecraft to allow operation in the event of failure dates back many years. In manned vehicles, the crew normally has had the capability to choose the backup, usually degraded in performance, but may have been incapable of making the decision in dynamic situations.

The NASA Space Shuttle, which had a design requirement to remain fully operational with one failure and to remain safe after two failures, probably reached the maximum in complexity in the approach to automatic and manual redundancy management. Basic to the design approach is the provision of four redundant primary flight computers to handle two computer hardware failures automatically and a fifth identical backup flight computer loaded with dissimilar software to be chosen in a manual switchover in case of a generic software error in the primary set. Three or more redundant channels of sensors are shared by all computers, which then command four parallel redundant channels to the effector servoelectronics. The primary flight computers automatically compare

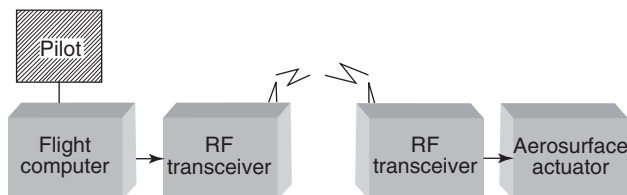


Figure 13. Fly-by-wireless concept.

and vote by “majority rule.” In addition, the aerosurface actuators, which have four channels at the secondary (pilot spool) stage, were designed to vote effectively by force fight, that is, three channels overcome the offending fourth channel by force summing three against one. A comprehensive discussion of these features can be found in Ref. 30.

The requirement of redundancy and automatic redundancy management in spacecraft depends on the need for vehicle survivability weighed against the cost of redundancy implementation. For expendable vehicles, it may be cheaper simply to build more spacecraft than to make them internally redundant. In manned spacecraft, loss of life is so intolerable in this country that safety must be maintained, regardless of cost.

Simulation

The sheer quantity of “black boxes” in a modern day GN&C system begs the question as to whether system end-to-end mathematical analysis is possible. Further, the time-varying, cascading signal flow from the multiple sensors to the flight computer to the various driver electronics for the motion effectors makes end-to-end analysis and solutions by conventional closed form methods virtually unachievable. Moreover, if the solution were tractable for a particular time and set of physical conditions and produced acceptable results from an operational point of view, the flight conditions could change in the next few moments, and the analysis would have to be repeated. The analytical approach commonly used in these situations is known as *computer simulation*, or simply *simulation*, where time is the primary independent variable.

Simulations are used in the conceptual design phase of a program and grow in complexity and fidelity as the program matures. Mathematical models of the functions of the different pieces of hardware evolve in complexity from simple first-order equations to higher and higher levels of fidelity. Desktop computers are adequate in the early phases to define and analyze single sensors, signal amplifiers, effectors, and units as complex as inertial platforms. As the program progresses, multiple hardware units are combined, vehicle characteristics and operational environments become better known in more detail, and the simulations are forced to move to larger complexes of computers, data library devices, and resultant data readout devices.

Additional complexity accompanies increased fidelity when the simulation is required to operate in a *real-time* mode. *Real time* means that a series of events simulated is calculated on the simulator in the same period of time as the real events would occur. Real-time simulation is used any time there is flight crew involvement or flight-like hardware is substituted for simulated hardware. For GN&C system-level simulations, the flight computer is usually one of the earliest units substituted for its modeled counterpart.

If the flight vehicle is a manned spacecraft, a crew station (cockpit) is usually added. Active crew displays and controls, either flight-like or simulated, are added to allow realistic crew participation for evaluating the GN&C system. If visual cues are required for this crew participation, “out-the-window” scenes may be included. The NASA Johnson Space Center Engineering Simulator that has

some two million lines of computer code is an example of these large man-in-the-loop simulations.

As the program matures, a simulator especially designed for flight crew and ground operations personnel training is usually constructed. This simulator-trainer typically has the highest fidelity of all simulations in the program. All onboard displays and controls are included, and system simulated failures may be introduced to enhance training. For manned spacecraft, out-the-window scenes are included for all phases of the intended vehicle mission, including prelaunch, launch, orbit transfers, midcourse deltaV's, rendezvous and docking with another spacecraft, deorbit, and landing.

Vital to the fidelity of simulations throughout a program are valid mathematical models for the components of the GN&C system, for effectors and other non-GN&C equipment, and for the environment in which the vehicle is supposed to be operating. Simple first-order equations are quite satisfactory early in the design phase, but as actual hardware is built, the fidelity desired increases, and the corresponding mathematical models become more complex. Similarly, equations of motion, environmental models for ambient physical conditions, aerodynamic data, ephemerides, and out-the-window scenes all become more intricate in the drive for fidelity as the program progresses and the design matures. This often results in large simulation complexes that are expensive to create and maintain, but they are indispensable for system development, anomaly investigations, and crew training.

Integration and Verification

The term *integration* is an overworked word in the aerospace business, but it is quite descriptive in GN&C systems development. Hardware piece parts and modules are *integrated* (assembled) into *line replaceable units* (LRUs) or "black boxes"² that are replaceable as the first level of maintenance on spacecraft. The LRUs are then *integrated* (combined) to form the whole GN&C system. A similar process is followed in developing flight computer software. As this process is followed, engineering tests are conducted, and design errors are uncovered and corrected. This successive *integration* is part of the development process and is the most demanding part of GN&C engineering.

Verification is the process of formal evaluation of the hardware and software and is done somewhat in parallel with the *integration* process. Four methods are generally used and are listed here in order of preference: *test*, *inspection*, *demonstration*, and *analysis*. These are defined as follows:

Test—the stimulation of the hardware and software under prescribed conditions and the responses measured and evaluated against specifications.

²The terms *black boxes* and *line replaceable units* (LRU's) have their heritage in the airplane world. Before the space era, most airplane electronics boxes were, and still are today, painted black. Line replaceable units are the level of replacement for maintenance on the airport *flight line* or ramp. A term that has come into use on the International Space Station program is *orbital replacement unit* (ORU), and as the name implies, the ORU is the usual replacement unit for maintenance during orbital flight around Earth.

- Inspection*—the visual examination of hardware, usually in static situations.
- Demonstration*—operation in a test-like environment but where responses usually cannot be measured and performance must be evaluated subjectively.
- Analysis*—mathematical and logical evaluation using mathematical models, drawings, flow charts, and photographs.

Formal hardware verification testing begins with the *acceptance test* at the GN&C system vendor’s plant. This test is conducted before the buyer accepts delivery to ensure that the equipment is functioning in accordance with the requirements of the contract between the vendor and the buyer.

Ideally, one or more LRUs are randomly selected from those already accepted for a subsequent *qualification test*. The qualification test is conducted at extremes in ambient conditions [pressure, temperature, vibration, input electrical power, electromagnetic interference (EMI), etc.] beyond the limits expected in normal operation. This test ensures that the hardware has comfortable operating margins.

The next level of testing is the *integrated system test*. In this test, the entire GN&C system is assembled in a flight-like configuration, and stimuli at the sensors are processed through the system to the outputs of the effector driver electronics. This is also sometimes called a *system end-to-end test*. Debate continues as to whether this test is an engineering development test or a formal verification test, but in reality it is usually both. Certainly, any design errors or generic manufacturing errors that are uncovered must be corrected. A critical part of this exercise is the testing of interfaces with non-GN&C equipment. The importance of this aspect is discussed in Ref. 31.

The final ground verification test is the *mission verification test*. The significant addition to the integrated system test configuration is a real-time simulation computer complex. This is used to “close the loop” (i.e., close the flight

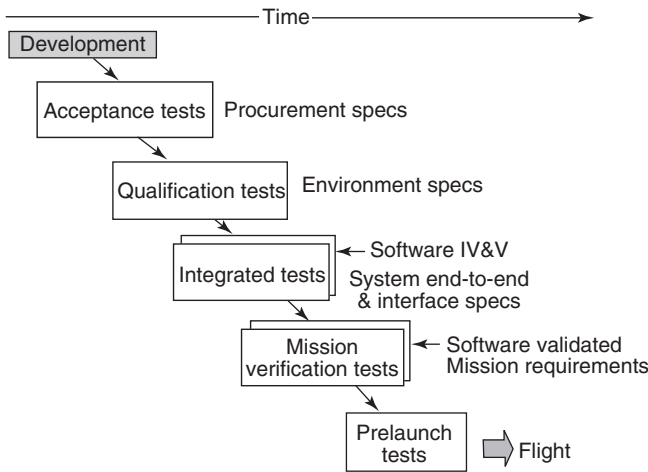


Figure 14. Integration and verification sequence.

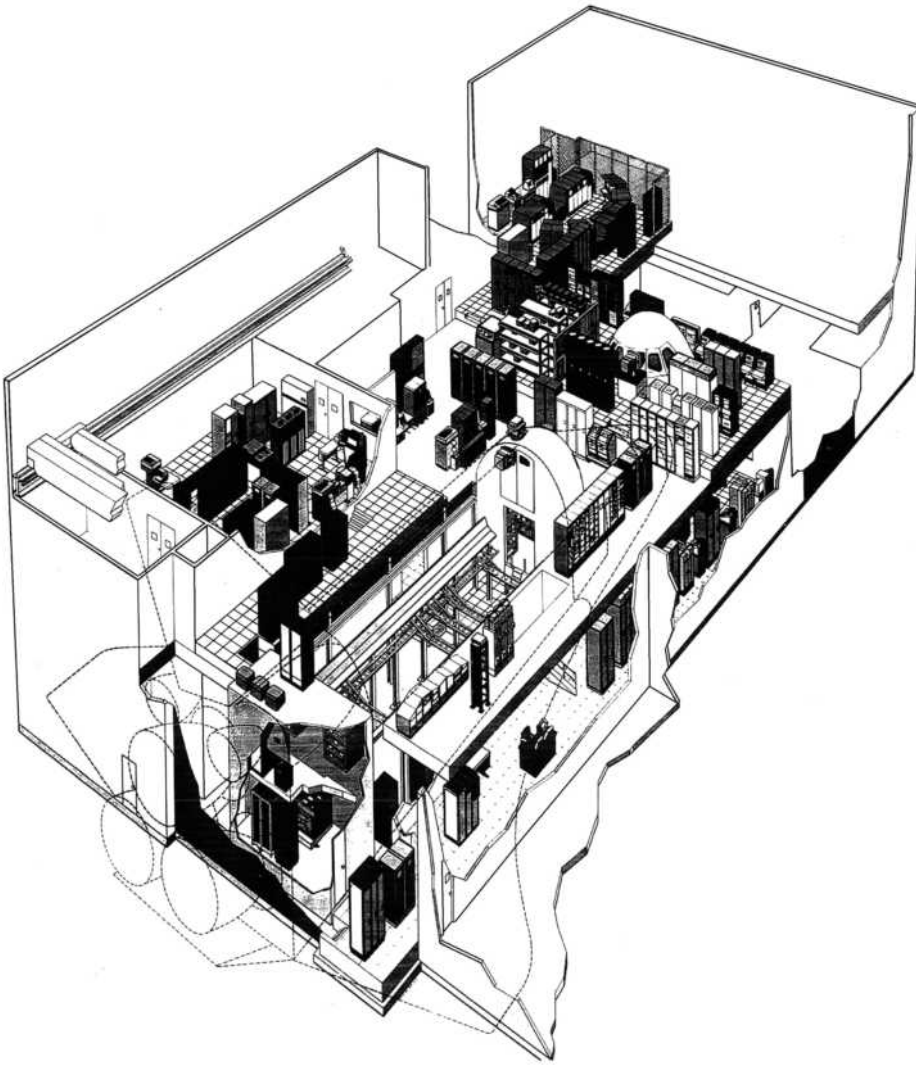


Figure 15. Shuttle Avionics Integration Laboratory (courtesy of NASA).

control loop using simulated vehicle dynamics) and to model missing flight hardware and the vehicle flight environment so that the entire mission, or significant portions, may be “flown” as realistically as practical short of actual flight. Flight computer software is used in the flight computer and is *validated* in the process. Successful completion of the mission verification test allows *certification for flight*. This sequence is shown pictorially in Fig. 14.

One of the most sophisticated facilities for the integrated system test and the mission verification test is the Shuttle Avionics Integration Laboratory at the NASA Johnson Space Center. An artist’s rendering of the facility is shown in Fig. 15. The avionics are installed in a high-fidelity, three-dimensional full-scale

arrangement with flight-like vehicle wire harnesses. The outline of the Orbiter is shown in dashed lines for orientation.

For manned spacecraft, one or more flight tests are usually required before the vehicle, including the GN&C system, is certified for normal operational use.

BIBLIOGRAPHY

1. Rauscher, M. *Introduction to Aeronautical Dynamics*. Wiley, New York, 1953.
2. Slater, J.M., and J.S. Ausman. Inertial and optical sensors. In C.T. Leondes (ed.), *Guidance and Control of Aerospace Vehicles*. McGraw-Hill, New York, 1965.
3. Sagnac, G. L'ether lumineux demontre par l'effet du vent relatif d'ether dans un interferometre en rotation uniforme. *Comptes Rendus d'Academie des Science Francais*, 95: 708–719 (1913).
4. Commission on Engineering and Technical Systems. *Technology for Small Spacecraft*. National Academy Press, Washington, DC, 1994.
5. Lefevre, H. *The Fiber-Optic Gyroscope*. Artech House, Boston, 1993.
6. Bryan, G.H. On the beats in the vibrations of a revolving cylinder or bell. *Proc. Cambridge Philos. Soc.* VII: 101 (1890).
7. Wright, D., and D. Bunke. The HRG as Applied to a satellite attitude reference system. In Guidance & Control 1994, Vol. 86, *Advances in the Astronautical Sciences*, R.D. Culp and D.R. Rausch (eds), *Proc. Annual Am. Astronaut. Soc. Rocky Mountain Guidance Control Conf.*, February 2–6, 1994, Keystone, CO.
8. Söderkvist, J. Micromachined gyroscopes. *Sensors and Actuators A* 43: 65–71 (1994).
9. Juneau, T., W.A. Clark, A.P. Pisano, and R.T. Howe. Micromachined rate gyroscopes. In *Microengineering for Aerospace Systems*. The Aerospace Press, El Segundo, CA, 1999.
10. NASA Manual ND-1021043, Apollo Command Module Block II Primary Guidance, Navigation, and Control System Manual, Rev. AA, NASA Contract 9-497, 10 March 1966.
11. Connelly, J., et al. MEMS-based GN&C sensors for micro/nano satellites, Guidance & Control 2000, Vol. 104, *Advances in Astronautical Sciences*, R.D. Culp and E. Dukes (ed.), *Proc. Annu. Am. Astronaut. Soc. Rocky Mountain Guidance Control Conf.*, February 2–6, 2000, Breckenridge, CO.
12. NASA Training Manual, National Space Transportation System Reference, Vol. 1, Systems and Facilities, June 1988.
13. Thompson, W.T. *Introduction to Space Dynamics*. Dover New York, 1986.
14. Ebinuma, T., R. Bishop, and E.G. Lightsey, Spacecraft rendezvous using GPS relative navigation. In *Spaceflight Mechanics 2001*, Vol. 108, Part 1, *Advances in the Astronautical Sciences*, *Proc. AAS/AIAA Space Flight Mech. Meet.*, February 11–15, 2001, Santa Barbara, CA.
15. DiPrinzio, D. and R.H. Tolson. Evaluation of GPS Position and Attitude Determination for Automated Rendezvous and Docking Missions, NASA Contractor Report 4614, Langley Research Center, July 1994.
16. Baker, R.M., and M.W. Makemson. *An Introduction to Astrodynamics*. Academic Press, New York, 1960.
17. Mueller, D.D. *Introduction to Elementary Astronautics*, unpublished class notes for a U.C.L.A. short course 1966.
18. Ball Aerospace and Technologies. *Aspect Camera Star Tracker*. Boulder, CO.
19. Thompson, W.T. Passive attitude control of satellite vehicles. In C.T. Leondes (ed.) *Guidance and Control of Aerospace Vehicles*. McGraw-Hill, New York, 1965.

20. Abzug, M.J. Active satellite attitude control. In C.T. Leondes (ed.) *Guidance and Control of Aerospace Vehicles*. McGraw-Hill, New York, 1965.
21. Kaplan, M.H. *Modern Spacecraft Dynamics and Control*. Wiley, New York, 1976.
22. Noton, M. *Spacecraft Navigation and Guidance*. Springer-Verlag, London, 1998. (Note: This monograph has free implementation software written in C++ available on the Internet.)
23. Wie, B. *Space Vehicle Dynamics and Control*, AIAA Education Series, Reston, VA, 1998.
24. Smith, G.H. Overview of aerospace vehicle computer applications. In C.T. Leondes (ed.), *Computers in the Guidance and Control of Aerospace Vehicles*, AGARDograph No. 158, February 1972.
25. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* March 1960.
26. Kalman, R.E., and R.S. Bucy. New results in linear filtering and prediction problems. *J. Basic Eng.* March 1961.
27. MIL-STD-1553B, Digital Time Division Command/Response Multiplex Data Bus, 1996 (replaces Rev. A dated 30 Apr 1975).
28. MIL-STD-1773, Fiber Optics Mechanization of an Aircraft Internal Time Division Command/Response Multiplex Data Bus, 02 October 1989.
29. NASA Tech Briefs, Vol. 24, No.2, Associated Business, New York, 2000, p. 8b.
30. Hanaway, J.F., and R.W. Moorehead. Space Shuttle Avionics System. NASA SP-504, 1989.
31. Euler, E.A., et al. The failures of the Mars Climate Orbiter and Mars Polar Lander: A perspective from the people involved. In *Guidance & Control 2001*, Vol. 107, *Advances in Astronautical Sciences*, R.D. Culp (ed.), *Proc. Ann. Am. Astronaut. Soc. Rocky Mountain Guidance Control Conf.*, January 31–February 4, 2001, Breckenridge, CO.

JON H. BROWN
Fort Worth, Texas

SPACELAB

Spacelab is the European contribution to the Post Apollo Program. U.S. President Nixon made an offer in 1969 to the eight member states of the European Space Agency to participate in the planned recoverable launcher area in developing the Space Shuttle and its payloads.

Spacelab is a reusable, multipurpose, modular laboratory and in the first 10 years of Space Shuttle operation, became its major payload. This payload remained attached to the Orbiter during the whole flight and maintained certain dependences. Spacelab was not a habitat, so that the Mission and Payload Specialists stayed in the Orbiter during take-off and landing and used it also as living quarters.

Spacelab was designed for a 10-year lifetime and/or 50 missions. Inside the laboratory, called the module, a shirtsleeve atmosphere for the astronauts was required; experiments on the outside had to be conducted in the open space on a pallet. Both module and pallet were accommodated inside the Shuttle payload bay and remained attached to the Shuttle during the mission (Fig. 1). This fact limited the time of a Spacelab mission to the time of a Shuttle flight.

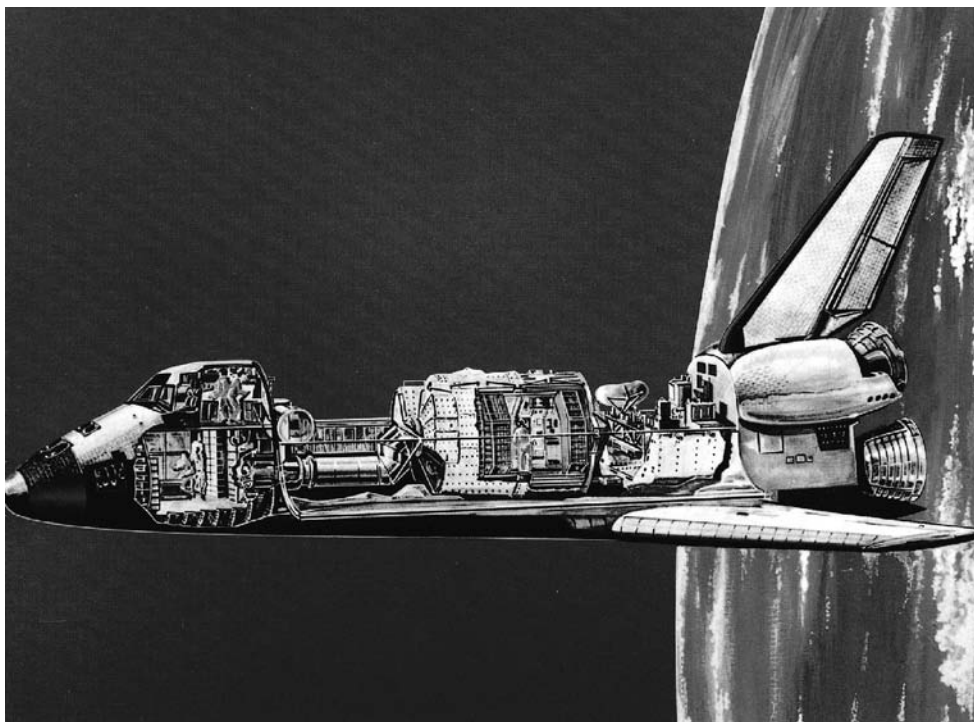


Figure 1. Spacelab. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

On 4 June 1974, the contract to design and build Spacelab was given by the European Space Agency (ESA) to an industrial consortium of companies from eight European countries under the leadership of the German space company ERNO Raumfahrttechnik in Bremen. The contract, won in a competition, was based on a modular design of the laboratory (Fig. 2) that applied state-of-the-art technology and experience in aircraft design to solutions of the turnaround requirements of frequent reuses.

For most of the time from 1974 to 1983, the Spacelab development ran parallel to the Shuttle development on a no exchange of funds basis to use common hardware and subsystems as much as possible.

The Modular Concept

The multipurpose objective of Spacelab led to a modular design to meet the needs of missions of very different natures for the module as well as for the pallet. To keep the cost down, the module and pallet elements were made a standard size, each 2.70 m. This allowed using standardized handling and transportation equipment and several combinations of module and pallet elements for the different missions. For pallet only modes, a special Igloo was

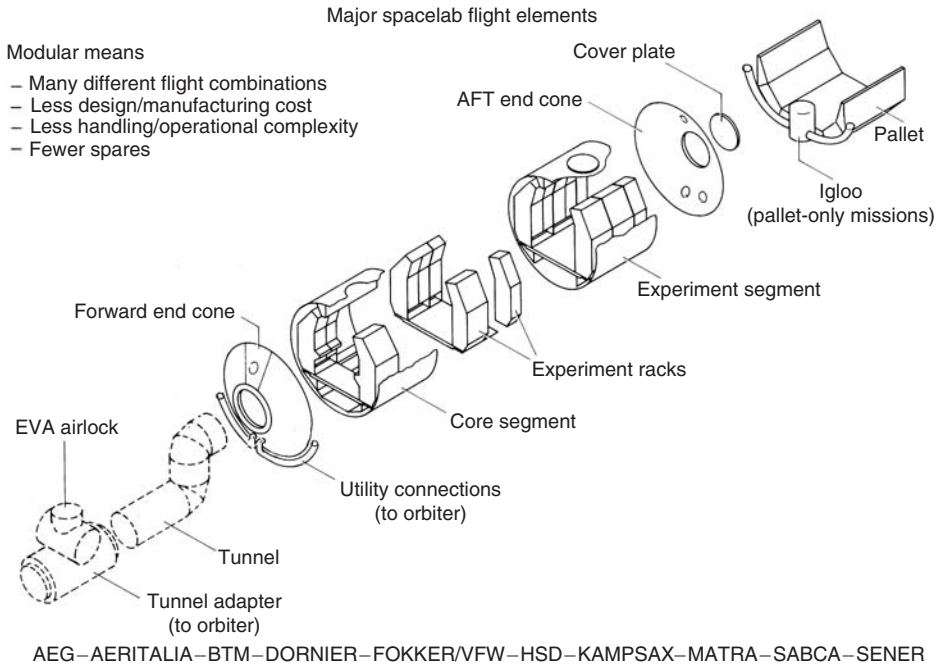


Figure 2. The modular concept Major Spacelab Flight elements.

developed to house the equipment that controlled the functions of the pallet and its payloads.

The Design Concept

The general shape and dimensions for this Shuttle payload were determined by the Shuttle payload bay. The Spacelab module has a cylindrical shape that is the diameter of the Shuttle payload bay. Although not needed in space, the cylinder contained a floor on which racks are mounted along the sidewalls.

The required 50 reuses and therefore, the corresponding 50 refurbishments on Earth demanded a floor/ceiling configuration to allow proper handling when under gravitational forces. This fact led to the solution to mount the racks on floor elements which could be rolled out of the cylinder like freight containers out of an airplane (Fig. 3). By this concept, the payload and experiments could be removed from the laboratory on their floor elements and shipped directly to their origin, and the new payload could be rolled in without any delay. By using this solution, the originally required 2-week turnaround time on the ground between missions could be achieved.

An aircraft structure design was selected for the pallet that allowed plain standardized surfaces for payload mounting and easy handling. A special instrument pointing system (IPS) was designed and built to be mounted on the

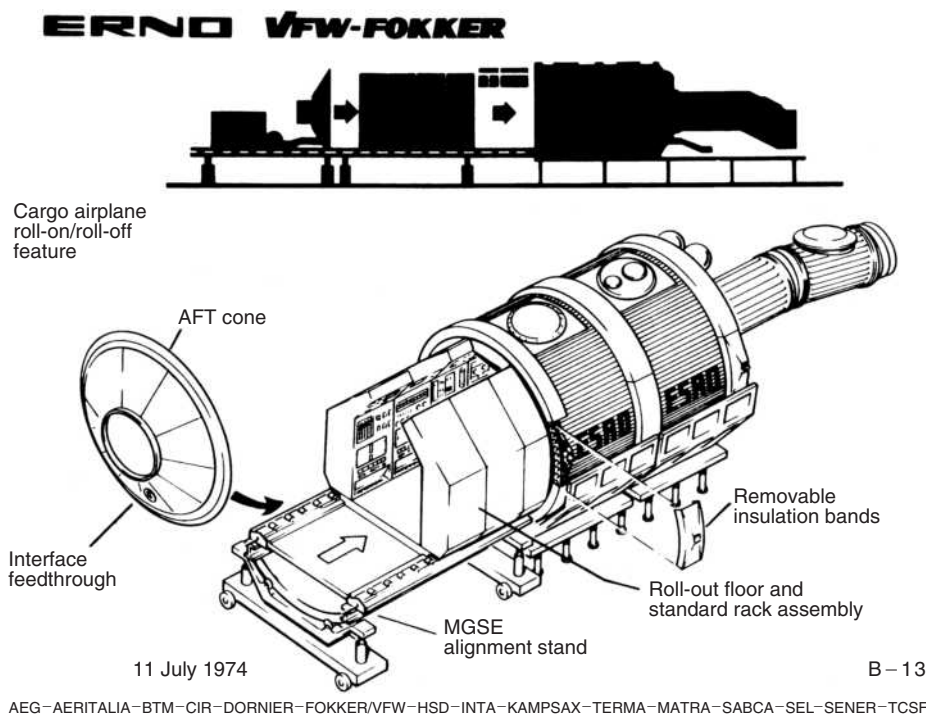


Figure 3. Payload integration—system level.

pallet to allow high accuracy pointing for payloads, thus eliminating errors from the Shuttle attitude control system.

An airlock on top of the module provided access of experiments to open space for direct exposure, and a window provided a view to the outside. Mission Specialists and Payload Specialists entered the laboratory through a tunnel from the lower aft flight deck of the Orbiter into Spacelab (Fig. 4). The 1.1-m diameter tunnel was not part of the European delivery program.

The Module Structure

The basic structural element of the Spacelab module was a cylindrical segment 2.70 m long and 4.06 m in diameter. This cylindrical shell was made of AL- 2219 alloy sheets that had a chemically milled inside waffle pattern and were welded at the seams. End flanges of forged aluminum alloy stabilized the cylinder and were used to join the segments to each other and to the end cones. Two of these segments, two end cones, internal floors, and several racks were part of the initial module structure's flight hardware deliveries. Cutouts in the shell for the window and the airlock required local strengthening to avoid distortion problems and to ensure pressure-tight mating surfaces (Fig. 5).

The floor could take loads up to 300 kg/m at the center and up to 500 kg/m at the outsides, and the racks could hold a load of 580 kg/m per side. The module

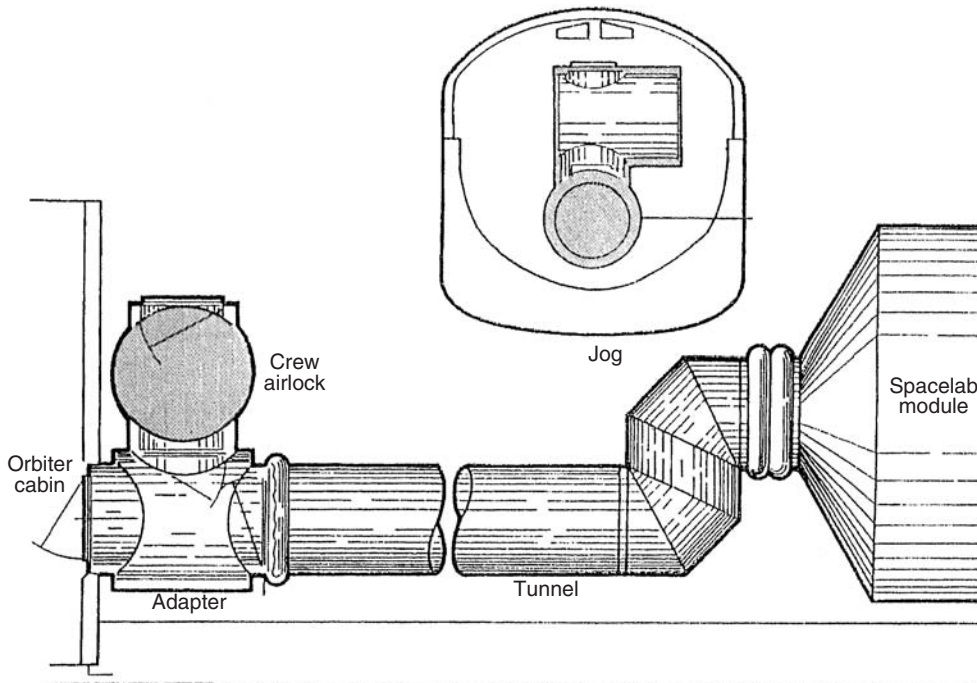


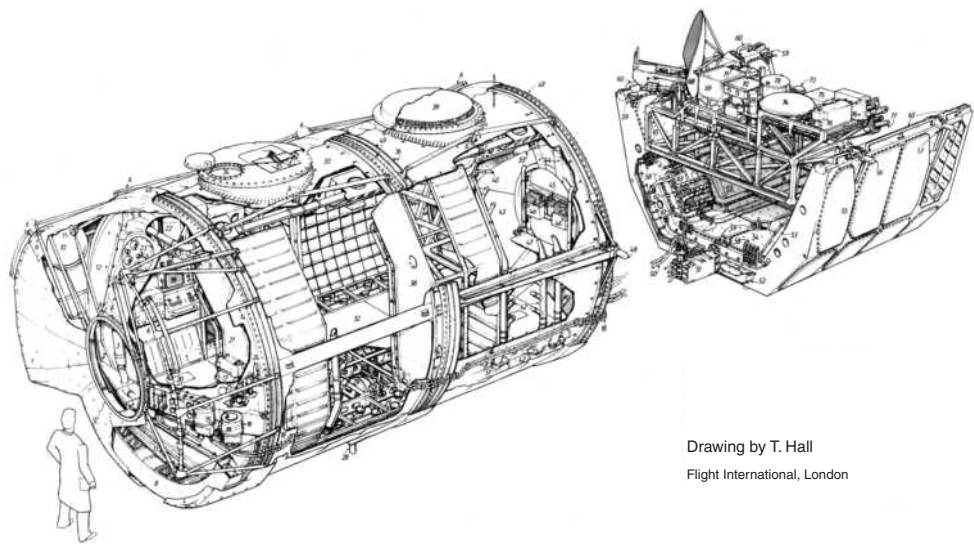
Figure 4. Offset concept for the transfer tunnel showing the jog required to permit centerline entrance to the Spacelab module.

and the pallet were supported in the Shuttle payload bay by trunnions and keel fittings which had to carry the weight of the loaded module on the ground and all loads imposed during launch and landing. A special surface finish had to be developed for these attachments to meet the very low friction coefficient to allow for twisting and bending of the Shuttle payload bay imposed by changing pressure, thermal loads, and landing impact. The module segment was attached by three fittings to the Orbiter payload bay, two trunnions and one keel fitting.

The subsystem equipment which was needed to run the laboratory was to be placed in two double racks at the forward end of the module and on a subfloor beneath the main floor. All other single or double racks were available for the payload. There were up to seven overhead storage containers in each segment attached to overhead supports mounted on the module ceiling, which also served as attachments for the racks.

The Pallet Structure

Like the module segments, the pallet segments were identical modular units 2.90 m long that could be mounted separately or together in pallet trains of two or three units or in combination with one or two module segments in the Shuttle payload bay. The basic U-shaped design allowed mounting experiments of up to 3000 kg on the inside or about 1100 kg/m equally distributed on standardized



Drawing by T. Hall
Flight International, London

Key			
Module			
1 Insulation blanket	13 Module/Orbiter lower feed-through plates (two)	27 Position for single rack	41 Overhead lights
2 Close-outs (skin/racks)	14 Insulation blanket supports	28 Keel-fitting	42 Avionics cooling-air ducts
3 Cabin air ducting (from subfloor)	15 Freon pump	29 Subfloor	43 Aft end-cone
4 RAAB	16 Water pump	30 Aluminium alloy module shell	44 Radial support structure
5 High data-rate recorder	17 Lithium hydroxide cartridge stowage	31 Electrical connectors for rack	45 Fire extinguisher (Halon)
6 Handrails	18 Freon lines	32 Floor of aluminium-skinned honeycomb sandwich (centre panel fixed, outer panels hinge up for access)	46 Portable oxygen equipment
7 Water/freon heat exchanger	19 Control-centre rack	33 Overhead duct channels	47 Foot restraint
8 Utility tray	20 Debris traps	34 Viewport	48 Module/Orbiter pickups (four)
9 Gaseous nitrogen supply line	21 Workbench rack	35 Nasa high-quality window	49 Module-segments joints, incorporating seals
10 Gaseous nitrogen tank	22 Stowage container (lower for access)	36 Fasteners for insulation blanket	Pallets
11 Temperature transducer	23 Upper module/orbiter feed-through plate	37 Rack fire-suppression system	50 Freon lines from module
12 Forward end-core	24 Gaseous nitrogen fill-valve bracket	38 Double rack	51 Pallet interface
	25 Gaseous nitrogen reducing valves (two-stage)	39 Experiment airlock	52 Cable ducts
	26 Position for double rack	40 Airlock controls	
	53 Cold plates	Experiments	
	54 Inner skin-panels	68 Synthetic aperture radar	
	55 Outer skin-panels	69 Solar spectrum	
	56 Pallet/Orbiter primary pickup	70 X-ray astronomy	
	57 Pallet/Orbiter stabilizer pickup	71 Solar constant	
	58 Connector support bracket	72 Charged-particle beam	
	59 Pallet herd-points	73 Advanced biostack	
	60 Handrails	74 Isotopic stack	
	61 Support systems remote aquisition unit (RAU)	75 Microorganisms	
	62 Experiment RAU (several)	76 Lyman Alpha	
	63 Experiment power distribution box	77 Waves	
	64 Pallet/bridge supports	78 Low energy electron-flux	
	65 Experiment-supporting bridge		
	66 Electrical junction box		
	67 Integrally-machined aluminium-alloy ribs		

Figure 5. Spacelab, Europe's space laboratory.

panels and on hard points that had an available volume of 33 m³ per pallet. Very thin aluminum faceplates were used on the panels to save weight. Four symmetrical support fittings were applied to suspend the pallet in the Orbiter payload bay (Fig. 6).

For the pallet-only missions, a special equipment container, the so-called Igloo, was developed. The primary structure of the Igloo was a cylinder-shaped

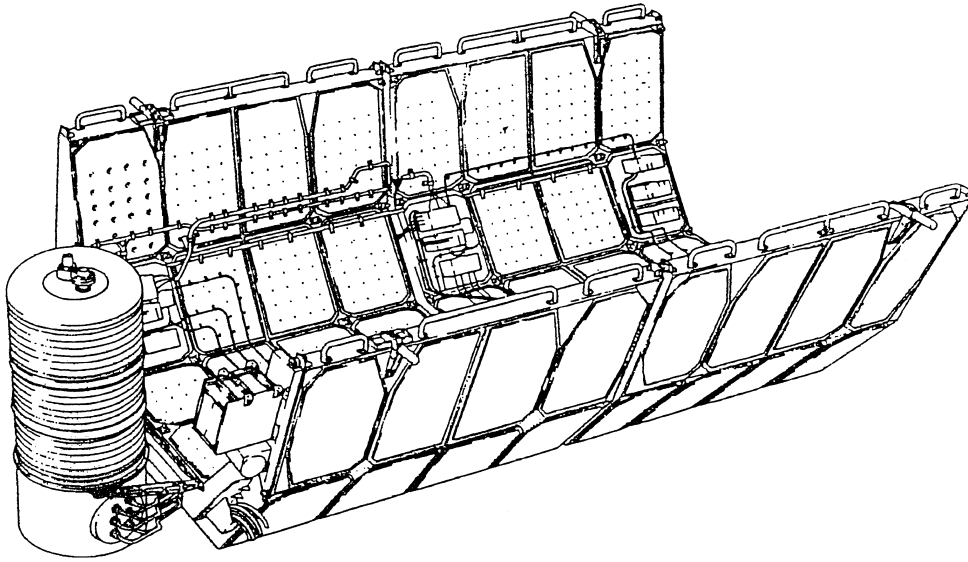


Figure 6. Pallet structure.

locally stiffened shell made of aluminum alloy forged rings. The lower end was closed, and the upper end was a mounting flange. The Igloo was attached directly to the forward end frame of the foremost pallet. The secondary internal structure was mounted on the primary structure and was finally closed by a cylindrical cover which was also a shell of aluminum alloy closed at the top and mated to the top flange of the primary structure. For the connection to the Orbiter, feed-through of utility lines was provided by penetrations of the Igloo. The thermal control of the Igloo was both active and passive; most equipment was cold-plated to the FreonTM system. The internal atmosphere was air, and a drying agent was included to avoid condensation after closing the cover. Nitrogen was added via a fill valve to ensure sufficient internal pressure during the mission. Overpressure safety protection was provided redundantly by a relief valve and a burst disk.

The Instrument Pointing System

One of the most interesting and ambitious subsystems was the instrument pointing system (IPS) for telescopes and sensors. The IPS provided three-axis attitude control and stabilization of payloads up to 2000 kg mounted on the pallet and exposed directly to open space. The pointing accuracy was ± 1 arcsecond for a payload that has a diameter up to two meters and a length up to 4 meters (Fig. 7). The design solution was an end-mounted approach in which the three gimbal systems would be mounted on the pallet providing support to a circular mounting frame to which the optical instrument would be attached. In zero gravity operation, the gimbal support structure would handle only the momentum of the instrument masses. During launch and landing, a payload gimbal separation mechanism would separate the IPS and the payload, a gimbal latch mechanism

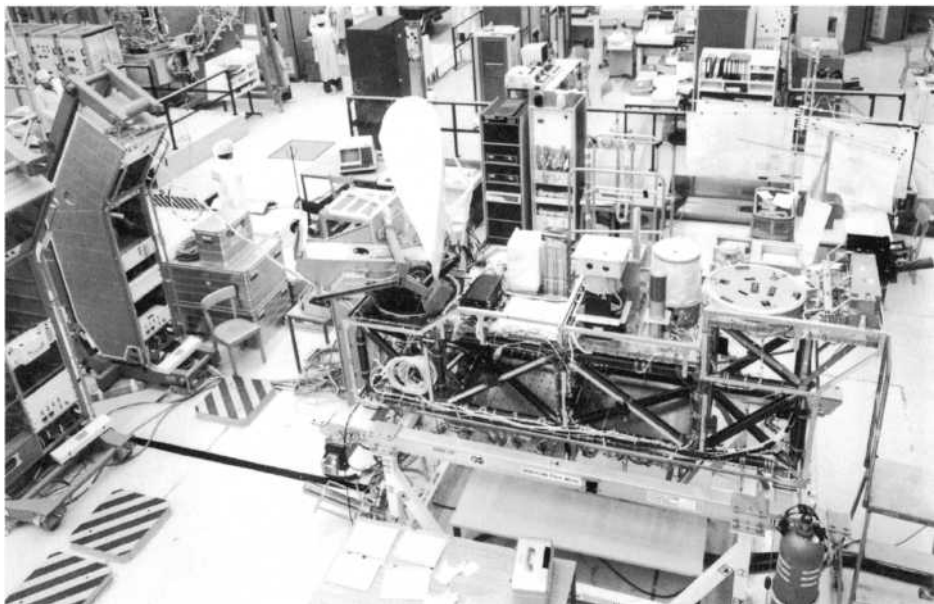
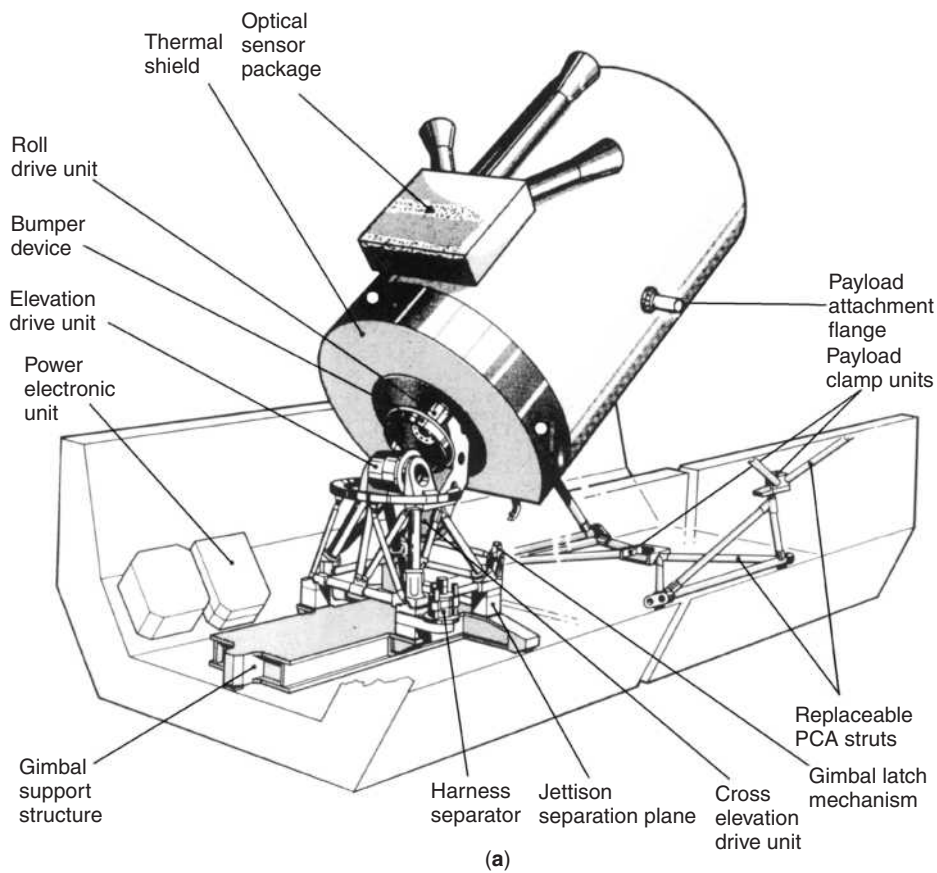


Figure 7. (a) Instrument pointing subsystem (b) Spacelab experiments testing.

would lock the gimbal system, and a payload clamp assembly would provide a three-point clamping of the payload.

The Environmental Control Subsystem

The environmental control subsystem (ECS) of Spacelab had to ensure the requirements of sea-level pressure and a shirtsleeve atmosphere for the crew and the experiments in the laboratory. It was agreed among the partners that the Orbiter would supply the oxygen and the Spacelab would provide its own nitrogen as well as the components to regulate and distribute the air supply throughout the module and remove excess moisture and carbon dioxide. A high-pressure nitrogen tank was mounted on the outside of the module. The CO₂ created by the breathing of the crew was absorbed by LiOH cartridges mounted on the subfloor underneath the main floor of the Spacelab cabin as well as by the two air fans and the contamination filters.

The heat generated by the module and the pallets was transferred to the Orbiter thermal control system. The minimum cabin temperature was supposed to be 19.5°C within a range of 18–27°C and the humidity between 30 and 70%. One circuit of air served the interior of the module cabin at a flow rate of 5–12 m/min, and a second circuit served the cooling of the equipment mounted in the racks. By passing through a heat exchanger, the heat load of the air was transferred to a flow of water that was pumped through an Orbiter payload heat exchanger to transfer the total Spacelab heat load finally to the externally mounted Orbiter radiators. The total capacity of this system was 5.8 kW.

A further component of the thermal control system was a FreonTM cooling circuit serving experiments mounted to cold plates on the pallet floor or lower sides. A total of eight such cold plates could be mounted on the pallet to accommodate experiments. Four thermal capacitors allowed storing of peak heat loads. As coolant, FreonTM 114 was used because water would freeze immediately when exposed to the outside environment of the pallets.

The Command and Data Management Subsystem

The most complicated subsystem of Spacelab is the command and data management subsystem (CDMS). It is an extension of the Orbiter telecommunication system which finally transmits data generated by the laboratory and its experiments which have been acquired by the CDMS and multiplexed in low-rate housekeeping and high-rate scientific data streams. NASA used two real-time data transfer systems. The Space Tracking and Data Network (STDN) transferred relatively low data rates of up to 192,000 bits/s from the Orbiter to the ground (downlink) and 72,000 bits/s from the ground to the Orbiter (uplink). Higher data rates were transmitted with the help of the Tracking and Data Relay Satellite System (TDRSS). This link allowed downlink data rates up to 50 million bits/s. On the ground, the data were received by ground stations in White Sands, New Mexico, via S-band or Ku-band through the geostationary satellites of the TDRSS. From there, they were relayed to the Payload Operations Control Center

at the NASA center in Clearlake near Houston and to the Spacelab Data Processing Facility at the Goddard Space Flight Center in Greenbelt, Maryland.

The CDMS had to provide mass memory storage for the central computers and for the periods of time when the Shuttle was out of sight of the TDRSS storage of high-rate digital data. A high-rate multiplexer ensured the handling of up to 16 channels of 16-megabit-per-second data.

Onboard computers, remote acquisition units that interfaced the functional laboratory equipment and the experimental equipment to the data bus, voice digitizer and intercom equipment, data display/keyboards, input/output and interconnecting stations, and a high-rate digital recorder were further elements of the CDMS.

In principle, the CDMS was divided into two elements: the data processing assembly and the high-rate data assembly. In the data processing assembly, two remote acquisition units separately served the Spacelab subsystems and the experiments. Each of these parts had a computer, an input/output unit, and a 1-megabit-per-second digital data bus which was routed through the laboratory. Three keyboards and displays and the mass memory served the two parts. This assembly was to acquire data; to distribute timing and commands via the data bus; and to interface with the Orbiter multiplexer/demultiplexer, the pulse code modulation unit, and the master timing unit. The high-rate data assembly consisted of the high-rate multiplexer, the high-rate recorder, and the demultiplexer and high-rate data recorders on the ground. This assembly acquired data directly from experiments and time division multiplexed these data into a composite data stream of up to 48 megabits per second, which could be transmitted to the ground by the Ku-band communication system. Besides the high-rate data, some low-speed data, as well as digitized voice and timing data, could be added into this composite data stream.

The Electrical Power Distribution Subsystem (EPDS)

Spacelab received its power from the Shuttle Orbiter. The power was produced by three fuel cells that converted oxygen and hydrogen into electrical power by a chemical reaction. In orbit, the power of one of these three fuel cells could be dedicated to Spacelab. This meant providing 7 kW of dc at 28 volts. The maximum power for Spacelab could be raised to 12 kW for 15 minutes once within a 3-hour time frame. Inverters (400 Hz) in the laboratory could convert parts of the energy into three-phase ac power at 115 and 200 volts. Built-in control and regulation circuits provided protection for inverters and consumers against overvoltage and overcurrent. During takeoff and landing, the available power for Spacelab was reduced to 1 kW because most of the experiments and subsystems were dormant in these mission phases. The total available quantity of oxygen and hydrogen in the Shuttle Orbiter was the factor limiting the total energy available for a mission. This limited a normal mission to about 300 kWh available for experiments in the Spacelab module configuration and 550 kWh in a Spacelab pallet-only configuration. Additional fuel cells in the Shuttle Orbiters "Columbia" and "Endeavour" allowed extending the mission's duration to 2 weeks as far as the power was concerned.

Basically, of the 7 kW power available for Spacelab, 5 kW was required by the laboratory subsystems and mission-dependent equipment, leaving only 2 kW for the experiments; 2 kW in the case of the pallet-only mode, only 5 kW remained for experiments. The main power conditioning, distribution, and control of the power through a feeder from the Orbiter took place in a power control box. This box included a shunt regulator to limit the main bus voltage to 32 volts and melting fuses to prevent short circuits on the feeders. A subsystem power distribution box distributed the dc and ac power into dedicated subsystem feeders.

The Follow-On Production (FOP)

The NASA/ESRO MOU of 1973 provided for NASA, to procure at least one further Spacelab no later than 2 years before the delivery of the first, provided that it met the agreed specifications and schedules and was reasonable in price. This agreement was realized by a procurement contract signed at the end of January 1980 and a preceding letter contract for the procurement of the essential long-lead items.

The Mechanical Ground Support Equipment (MGSE)

Parallel to the development of the Spacelab module and pallet and their auxiliary subsystems, sophisticated ground support equipment had to be developed that mirrored the dimensions and accuracy requirements of the flight hardware. Besides the integration process in its different phases, ground handling included transport on the ground and in the air.

The modular design of Spacelab required a mechanical ground support equipment that allowed mating two modules and their two end cones on the ground to simulate the real flight configuration with respect to the accurate integration and verification of the O-ring seals. It also required mating a three-pallet train.

In addition to the integration equipment, the modules needed special transport equipment in which each module could be placed like a flat can, because its dimensions did not allow transport on the road in a normal upside position. Also, the eight double racks and four single racks, which were to hold avionics equipment and the experiments, had to be supported for integration, handling, and transport, including their floor sections onto which they were mounted.

Besides the mechanical support, access equipment was needed such as workstands, scaffolding, access kits, floor covers, soft covers, desiccants, and transport tie-downs. For simulation and test on the ground, the gaseous nitrogen system, as well as the fluid loop of the active thermal control system and the environmental control life support system, had to be provided.

Finally, checkout equipment for weight and balance, leak checks, and verification of the environmental control life support elements had to be provided.

The Electrical Ground Support Equipment (EGSE)

The parallel development of the Shuttle and Spacelab caused interface problems for the laboratory design because open issues or changes in the Shuttle development had a direct impact on the dependent payload, Spacelab. In particular, this problem existed for the electrical ground support equipment, which was to simulate the Shuttle on the ground during Spacelab integration, test, verification, and qualification. The electrical ground support equipment served for test and ground checkout during integration of the laboratory and during the integration of the payload into Spacelab. These tests and checkouts were to be conducted in a manual and in an automatic mode. Serving this purpose was Automatic Test Equipment (ATE), which was composed of a three-station console with keyboards and CTR displays, a CII MITRA 125 computer with related peripherals, recording and timing equipment, measurement components, stimulus generators, and interface equipment.

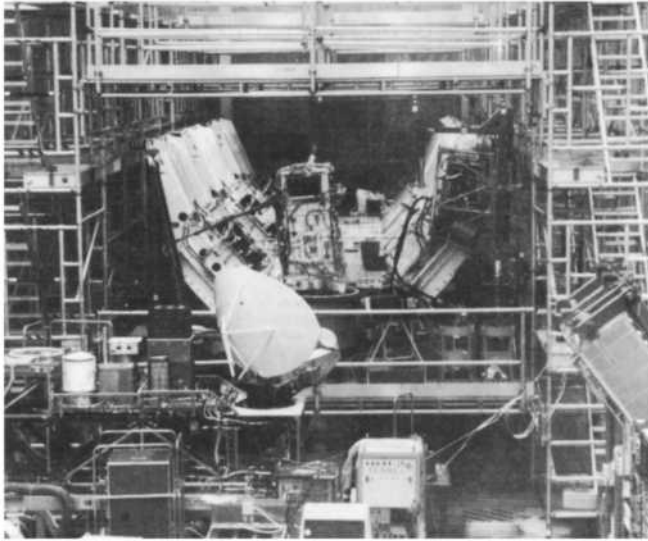
Further elements of the electrical ground support equipment were the ground power unit to simulate the Orbiter power supply, an Orbiter interface adapter to simulate the Orbiter power and signal interface, an experiment segment/pallet simulator to simulate power loads for the Spacelab electrical power system and signals for the command and data management system, and an experiment subsystem simulator to simulate experiment power loads and data interfaces.

Two complete sets of this electrical ground support system were developed and used first in Bremen for Spacelab development and later in the Kennedy Space Center (KSC) O and C Building to support mission preparation and payload integration.

The Flight Hardware

Although not contained in the original flight hardware delivery program, ESA and NASA agreed to use the Spacelab engineering model pallet structures for orbital flight tests (OFT). These pallets were kept very simple, equipped only with an Orbiter FreonTM pump, cold plates, a power control box, and a flexible module for command purposes. With this additional agreement about the European hardware, the first Spacelab OFT-pallet arrived at NASA's Kennedy Space Center in Florida on 4 December 1978. The pallet, packed in a special container, was shipped by sea in mid-November from Bremen to the Cape via Savannah, Georgia, and the Interstate Coastal Waterway. The second OFT-pallet arrived at KSC on 22 April 1979. They became part of the preoperational Orbiter payload, intended to collect data prior to the first operational flight of Spacelab.

On 21 December 1981 the two segments of the long module were flown out of the Hanover airport because the runway was appropriately long for the USAF C-5 "Galaxy" nonstop flight directly to the Kennedy Space Center in Florida. The distance from Bremen to Hanover was covered on the road. The Igloo and three pallets followed on 28 July 1982. The follow-on production (FOP) flight hardware was delivered to KSC on 27 July 1984 (Fig. 8).



(a)

Initial SL configuration	Pallets for shuttle orbital flight tests (OFT)	Engineering model (manned ground version)	Flight unit I (manned flight configuration)	Flight unit II (unmanned flight configuration)
Further possible SL configurations				
SL hardware delivered	2 OFT pallets plus related MGSE <ul style="list-style-type: none"> • 4 Dec. 1978 • 22 April 1979 	Long module, 3 pallets plus related MGSE, software documentation <ul style="list-style-type: none"> • 5/8/13 Dec. 1980 • 27 July 1981 EGSE set 1, servicers, software, MGSE, spares allowing operation of EM functionally and checkout of operational procedures	Long module, 2 pallets plus MGSE, unit testers, servicers, airlock (EM & FU) viewport adapter, assembly, spares, configuration and mission dependent items, loose parts, EM ship shorts, software documentation <ul style="list-style-type: none"> • 13 Nov. 81, 11 & 21 DEC. 81 	Igloo, 3 pallets plus EGSE set 2, MGSE, software, documentation <ul style="list-style-type: none"> • Planned for June 1982
SL hardware use in space shuttle flights	<ul style="list-style-type: none"> • 1 pallet used in STS 2, Nov. 12 to 14, 1981, carrying OST A 1 • 1 pallet for use during STS 3 planned for 22 March 1982, carrying OSS-1 	This model is the spacelab ground version used for checkout, training and integration functions with the space shuttle	According to NASA's flight manifest (Dec 1981) for the first 72 shuttle missions, the following use of spacelab is planned <ul style="list-style-type: none"> • Sep 1983, Spacelab 1 SRS-9 *long module + 1 pallet • Sep. 1984, Spacelab 3 STS-19 *long module • May 1985, Spacelab D1 STS-28 *long module • Sep. 1985, Spacelab 4 STS-33 *long module • Aug. 1986, Spacelab 8 STS-48 *long module + 1 pallet • Jan. 1987, Spacelab 6 STS-54 *short module + 3 pallet • May 1987, Spacelab 10 STS-58 *long module + 1 pallet • Sep. 1987, Spacelab J STS-63 *long module + 1 pallet 	<ul style="list-style-type: none"> • Nov. 1984 : Spacelab 2 STS 21 : Igloo + 3 pallets • Mar. 1987 : Spacelab D-4 STS 56 : Igloo + 4 pallets

(b)

Figure 8. (a) Spacelab Integration Hall at ERNO. Bremen Assembly of the Spacelab Flight Unit II, comprising Pallets plus Igloo. The Igloo (secondary structure in tilted position and without cover) is attached to the pallets, which are partly equipped with cold plates and subsystem equipment. (b) Spacelab hardware content/deliveries/configurations and use in pre- and operational Space Shuttle flights.

Schedule and Cost

From fall 1969 to summer 1972, the ESA and the European space industry investigated their possible contribution to the U.S. Post Apollo Program proposed by President Nixon. The direct participation in the Orbiter development and the development of a Space Tug by the Europeans was not accepted, and in 1973, ESA and NASA signed the agreement for Spacelab. After competitive study phases A and B, the development proposals were submitted to ESA by two industrial teams on 14 April 1974. The final contract was awarded to the ERNO team of Bremen on 5 June 1974, which was the Authorization To Proceed (ATP). With a great number of industrial and governmental partners from both sides of the Atlantic involved, the following milestones were achieved:

- 24 June 1974 Kick off
- 11 November 1974 Preliminary requirements review (PRR)
- 19 June 1975 Systems requirements review (SRR)
- 29 September 1975 Contract signature ESA/ERNO in Paris
- 02 July 1976 Preliminary design review (PDR)
- 03 March 1978 Critical design review (CDR)
- 04 December 1978 Delivery of the first OFT flight pallet to KSC
- 28 November 1980 Spacelab engineering model rollout
- 30 November 1981 Spacelab flight module (FU-1) roll out
- 11 December 1981 Spacelab flight module delivered to KSC
- 05 February 1982 Spacelab (FU-1) acceptance at KSC
- 08 July 1982 Spacelab flight pallets (FU 2) delivered
- 13 January 1983 Spacelab design certification review at NASA HQ
- 18 November 1983 Final flight readiness review at NASA HQ
- 28 November 1983 First Spacelab flight with STS 9 "Columbia"
- 27 July 1984 Follow-on production unit delivery

More than 9 years after development began, Spacelab conducted its first flight. Although 3 years longer than originally assumed, this duration did not delay the Shuttle schedule. The launch date of STS-9 was well met by the completion of the development of the laboratory and by delivery of the first Spacelab payload (FSLP).

The participation of nine European countries in the Spacelab program meant that the development work was executed by companies that had nine different currencies. ESA used, therefore, a neutral artificial currency for this project; the accounting unit (AU) was a kind of provisional EURO, which corresponded to the U.S. dollar with a slightly changing rate of exchange over time. The baseline cost of 193 MAU (million accounting units) in 1974 increased by changes in the scope of the contract, cost overruns, and a 103% cost escalation during the 9 years to 579 MAU. Not including the escalation, the overall program cost was below 140% of the agreed cost for this first European manned space-flight program. By the end of the program, Germany had contributed 56.79%,

Italy 11.34%, France 10.42%, the United Kingdom 6.67%, Belgium 5.29%, Spain 3.56%, the Netherlands 2.50%, Denmark 2.07%, Switzerland 1.07%, and Austria 0.29%.

The Missions

Spacelab was designed for a 10-year operational lifetime and/or 50 reflights. The laboratory exceeded its design lifetime by more than 4 years, but it did not achieve 50 flights. The first pallet flight occurred in November 1981, and the last in July 2001, a time span of almost 20 years. The first module flight took place in November 1983, and the last in April 1998, a time span of almost 15 years.

In total, 15 Shuttle flights carried a Spacelab module, 6 flights carried pallets plus an Igloo, and 10 flights carried pallets only (Tables 1 and 2). AN OFT pallet was used in the Orbiter “Columbia” to accommodate the Shuttle imaging radar SIR-A as early as the second flight. The third Shuttle flight STS-3 also contained a pallet in the payload bay.

The first flight mission of Spacelab SL-01 was the Shuttle flight STS-9, and the Orbiter “Columbia” was launched on 28 November 1983 from Kennedy Space Center Pad B at 332:16:00:00 GMT. The Spacelab consisted of a long module and a pallet whose total mass was 15,265 kg (33,252 lb), both up and down. The flight was scheduled for 9 days. However, NASA and ESA decided at midtime to prolong it by 1 day because the usage of the resources on board allowed such an extension. So the mission was completed when the Orbiter “Columbia” landed at Edwards Airforce Base in California on 8 December 1983 at 343:23:47:00, accomplishing a mission duration of 11 days, 7 hours, 47 minutes. In that time, the Orbiter rounded the globe 166 times and covered a distance of about 7.5 million kilometers at an inclination of 57° about the equator (Figs. 9 and 10).

The Spacelab mass was 12,780 kg (28,185 lb), of which 8145 kg (18,135 lb) were for the module and 3386 kg (7449 lb) for the pallet. The mass of the experiments and associated equipment was 3982 kg (8802 lb) and for the verification flight instrumentation 856 kg (2123 lb).

Spacelab 1 was a joint mission of ESA and NASA. Each agency sponsored about half of the scientific payload. In the United States, the Marshall Space Flight Center was assigned responsibility for the NASA sponsored portion of the payload. In Europe, the responsibility for the ESA sponsored portion of the payload, referred to as the first spacelab payload (FSLP), was entrusted to a technical management team known as Spacelab Payload Integration and Coordination in Europe (SPICE). This team was set up by ESA in 1976 at the German Aerospace Research Establishment (DFVLR), Cologne-Porz.

Spacelab 1 was a multidisciplinary mission of more than 70 experiments in five areas of scientific research: astronomy and solar physics, space plasma physics, atmospheric physics and Earth observations, life sciences, and material science. In total, 38 different facilities for these experiments existed. Of the 38 experiments, 16 required to conduct investigations were situated on the pallet and 20 in the module. Two of the 38 experiments had components both on the pallet and in the module.

Table 1. **Spacelab Module Missions**

STS carrier	Launch date & duration	Orbit incl. & alt.	Mission	Configuration	Discipline	European user participation		European astronaut
						Major	Partial	
STS 9 Columbia	28 Nov. 83 10 days	57° 250 km	SL-01 FSLP	LM + IP	Multidiscipl.	X		U. Merbold
STS 51B	29 Apr. 85	57°	SL-03	LM + MPESS	Mat. Science			
Challenger	7 days	360 km						
STS 51F	29 Jul. 85	50°	SL-02	IG + 3P + IPS	Solar Astron.		X	
Challenger	8 days	320 km						
STS 61A	30 Oct. 85	57°	SL-D1	LM + MPESS	Mat. Science	X		R. Furrer E. Messerschmid W. Ockels
Challenger	7 days	330 km						
STS 35	2 Dec. 90	28°	ASTRO-1	IG + 2P + IPS	Life Science Astronomy			
Columbia	9 days	350 km						
STS 40	5 Jun. 91	39°	SLS-01	LM	Life Science			
Columbia	9 days	300 km						
STS 42	22 Jan. 92	57°	IML-01	LM	Mat. Science		X	U. Merbold
Discovery	8 days	300 km			Life Science			
STS 45	24 Mar. 92	57°	ATLAS-1	IG + 2P	Atm. Physics		X	D. Frimout
Atlantis	9 days	300 km			Solar Astron.			
STS 50	25 Jun. 92	28°	USML-01	LM + EDO	Mat. Science			
Columbia	14 days	300 km						
STS 47	12 Sep. 92	57°	SL-J	LM	Mat. Science			
Endeavour	8 days	300 km			Life Science			
STS 56	8 Apr. 93	57°	ATLAS-2	1G + 1P	Atm. Physics		X	
Discovery	9 days	300 km						
STS 55	26 Apr. 93	28°	SL-D2	LM + USS	Multidiscipl.	X		M. Schlegel U. Walter

Table 2. Spacelab Pallet Missions

STS carrier	Launch data and duration	Spacelab pallet and purpose
STS-2 Columbia	12 Nov. 81 2 days	1 Pallet/OSTA-01-Shuttle Imaging Radar SIR-A
STS-3 Columbia	22 Mar. 82 8 days	1 Pallet/OSS-01-NASA Office of Space Science
STS-41G Challenger	5 Oct. 84 8 days	1 Pallet/OSTA-03-Photographic and radar images of Earth
STS-51A	8 Nov. 84	2 Pallets/Retrieval of Palapa and Westar communication satellites
Discovery STS-39	8 days 28 Apr. 91	1 Pallet/AFP-675-Air Force Program 675
Discovery STS-46	8 days 31 Jul. 92	1 Pallet/TSS-01-Tethered satellite
Atlantis STS-59	8 days 9 Apr. 94	1 Pallet/SRL-01-Space Radar Laboratory
Endeavour STS-64	11 days 9 Sept. 94	1 Pallet/LITE-Lidar In-Space Technology
Discovery STS-68	11 days 30 Sept. 94	1 Pallet/SRL-02-Space Radar Laboratory
Endeavour STS-75	11 days 22 Feb. 96	1 Pallet/TSS-1R-Tethered satellite reflight
Columbia	16 days	

Some of the experiments on the pallet and in the module operated automatically; others were operated from the ground or by the crew remotely through the computer or by controls located on the instrument front panels. Other experiments in the module were operated directly by the crew.

Of the 38 experiments, 13 were sponsored by NASA, and 25 by ESA. Of the 13 NASA experiments, 5 were located on the pallet and 8 in the module. Of the 25 European experiments, 11 were situated on the pallet and 14 in the module. The 38 experiments on board Spacelab 1 were selected from more than 400 proposals solicited by NASA and ESA in 1976. The STS-9 Shuttle/Spacelab crew consisted of six astronauts, five Americans and with Ulf Merbold, one German. The Commander was the experienced 52-year-old John Young on his sixth space flight. The crew was divided into two teams, a red team and a blue team, both to work in space in 12-hour duty cycle shifts:

John W. Young Commander (CDR) Red Team

Brewster H. Shaw Pilot (PLT) Blue Team

Dr. Owen K. Garriott Mission Specialist (MS 1) Red Team

Dr. Robert A. R. Parker Mission Specialist (MS 2) Blue Team

Dr. Byron Lichtenberg Payload Specialist (PS 1) Red Team

Dr. Ulf Merbold Payload Specialist (PS 2) Blue Team

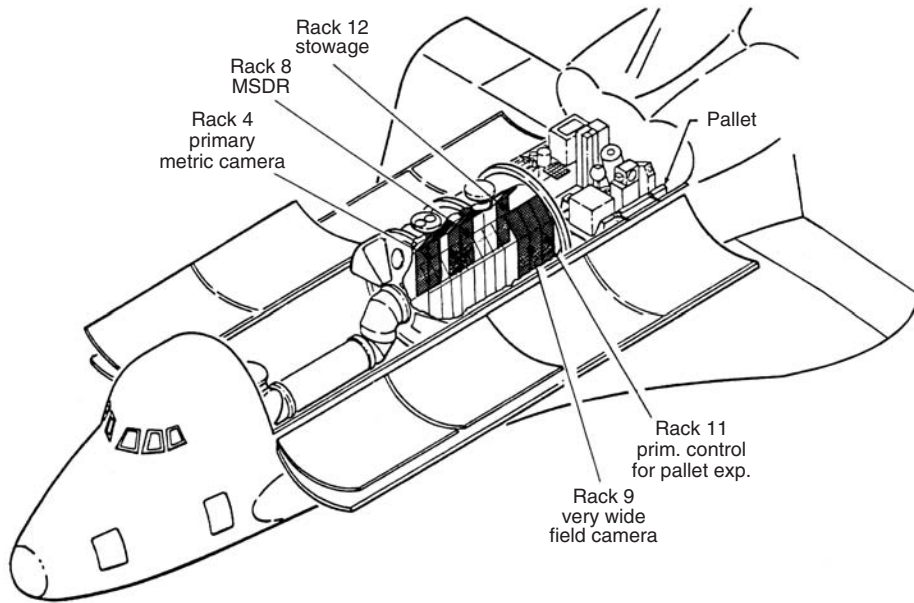


Figure 9. SL-1 configuration.

After the first Spacelab module flight SL-01, a second Spacelab mission, SL-02, with a pallet-only mode, including the Igloo and an instrument pointing system (IPS), was launched on 29 July 1985 and flown for 8 days. This mission specialized in solar astronomy. Both missions were joint NASA/ESA missions and were considered the conclusion of the development program.

After those missions, several other missions used Spacelab, both U.S. missions and international missions. With D-1 and D-2, Germany equipped two

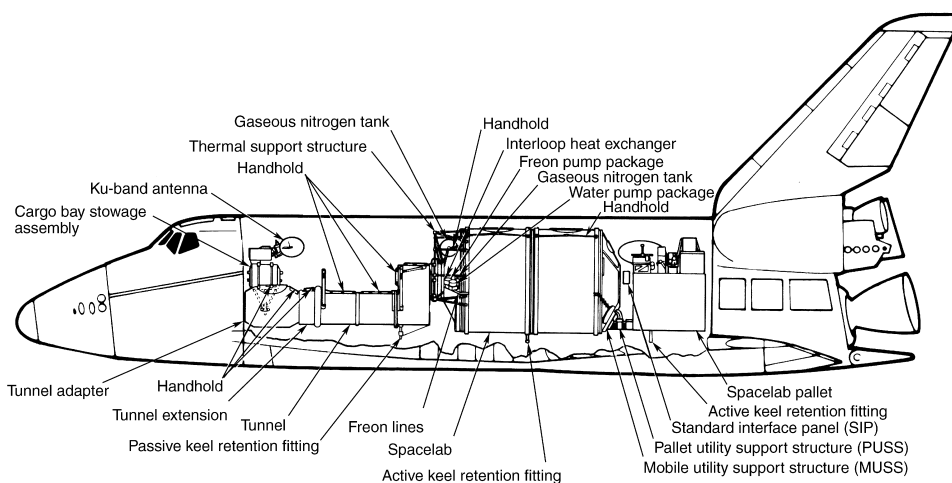


Figure 10. Side view.

multidisciplinary Spacelab missions. Spacelab D-1 was launched on 30 October 1985 and was very successful. Spacelab D-2 was planned after the tragic “Challenger” mission in early 1986 and was delayed accordingly for more than 4 years until 26 April 1993.

On 17 April 1998, the last Spacelab mission with a module took off into space with the Orbiter “Columbia” for a 16-day flight on a life science mission called NEUROLAB. This was 14 years and 5 months after its maiden flight in November 1983.

Lessons Learned

The Spacelab program was NASA’s largest international cooperation program, because it involved with ESA on the European side, nine countries and its major payload was developed, parallel in time to the Space Shuttle, a unique situation with respect to interdependences and interfaces. And for this situation, the partners had agreed to a “no exchange of funds” funding and use of common hardware as much as possible. High hopes were attached to the Shuttle as the first recoverable launch system in the Post Apollo Program and its major payload Spacelab. For the European partner, however, Spacelab was the first manned space project ever.

Under these circumstances, the program was a full success. On the European side, the selection of an industrial prime contractor for the first time for a project of this size was a solution, which proved to be very helpful, because most of the coordination effort between companies from nine European countries was executed within the consortium under the leadership of the prime contractor.

The employment of U.S. industrial consultants in those areas that lacked experience in Europe, especially in the aspects of manned space flight, proved to be a very supportive measure. Representatives on all levels in the governmental organizations and in industry helped to avoid misunderstandings and served as on the spot mediators and real-time informants. A very tight network of clearly defined periodic meetings, working groups, and committees on all levels in government and industry kept all parties involved, provided the information to the need-to-know, and challenged the responsibilities in the decision-making process.

The intentional application of state-of-the-art technology for the Spacelab development reduced the risk and was an important factor in keeping the schedule and cost plan that had so many partners of different experience involved. The frequent usage of the laboratory for very different missions with short turnaround times required well-known technologies and easy access to spare parts. The access to Electric/Electronic Equipment (EEE) parts during a period of more than 9 years of development and 15 years of operation was, however, a problem. During this period of more than 24 years, the fast advance of technology in the EEE parts industry caused problems of an undisturbed supply of parts for Spacelab.

The availability of only one prototype of Spacelab, EM-1, in the program for cost saving reasons, resulted in the fact that there was no prototype in Europe to follow the operational program after EM-1 was delivered to the Marshall Space Flight Center in Huntsville, Alabama, on 28 November 1980.

After the infrastructural system of Shuttle/Spacelab was completed, the partners, NASA and ESA, and the participating countries in Europe did not continue their joint effort for using of Spacelab as the public would have expected. There was no agreed upon organization for evaluating the results achieved in the laboratory and their importance for science, technology, and application. Attention was directed too soon to the next generation of a manned space facility, the Space Station. The limited mission time of the Shuttle-attached Spacelab was one of the reasons for this development, although it took more than 10 years before the decision for the next generation of manned orbital infrastructure was made.

Fortunately, much of the good Spacelab experience could be saved and flow into the even greater challenge of a permanent, manned station in low Earth orbit launched with the Space Shuttle. Again, a module will be supplied by Europe incorporating Spacelab experience and spirit.

BIBLIOGRAPHY

1. Lord, D.R. *Spacelab. An International Success Story*. NASA Scientific and Technical Information Division, Washington, DC, 1987.
2. Messerschmid, E., R. Bertrand, and F. Pohlemann. *Raumstationen, Systeme und Nutzung*. Springer-Verlag, Berlin, Heidelberg, New York, 1997.
3. Buedeler, W., and S. Karamanolis. *Spacelab, Europas Labor im Weltraum*. Wilhelm Goldmann Verlag, Muenchen, 1976.
4. Von ERNO bis Asrtrium, Wir zeigen die Vergangenheit und sprechen von der Zukunft Schriftenreihe der Raumfahrtshistorischen Archivs Bremen e. V. Heft 1. Stedinger Verlag, Lemwerder, 2001.
5. Sebesta, L. *Spacelab in Context*. European Space Agency, Paris, France, 1997.
6. Longdon, N. *Spacelab Data Book*. European Space Agency, Paris, France, 1983.

HANS E.W. HOFFMANN
ORBCOMM LLC
Dulles, Virginia

SPUTNIK 1: THE FIRST ARTIFICIAL EARTH SATELLITE

Humanity's leap into space was one of the greatest scientific achievements of the twentieth century. However, one of the ironies of history is that this great scientific and engineering achievement was largely facilitated by the Cold War between the two superpowers, the United States and the Soviet Union.

In the late 1950s, the United States held a significant advantage over the Soviet Union in number of nuclear warheads and capability to eliminate the most important strategic facilities on its enemy's territory. Although the Soviet Union already had nuclear weapons, it did not have second-strike capability because it

did not have any delivery vehicles capable of delivering warheads to U.S. territory. The senior political leadership of the Soviet Union assigned its missile science and industry a task of “special governmental importance”—development of a ballistic missile that could render the United States vulnerable.

The early experimental intercontinental ballistic missiles developed by Soviet scientists and engineers were used to place the first artificial Earth satellites into orbit. Thus, the history behind the development and launching of the first satellite into orbit is primarily the story of a missile that became a launch vehicle for a satellite rather than a missile carrying a weapon of mass destruction.

As long ago as February 1953, a government decree assigned NII-88, the lead institute for missile technology, the task of performing “theoretical and experimental research related to developing a two-stage ballistic missile and studying the future applications for a missile with a range of 7000–8000 km.” This task was led by Sergei Korolev, who at that time was the Chief Designer of OKB-1 (at that time part of NII-88). Korolev, for his part, was head of the Chief Designer’s Council, which included the following Chief Designers: Valentin Glushko for rocket motors, Nikolai Pilyugin for trajectory control systems, Vladimir Barmin for ground launch systems, Mikhail Ryazanskii for radio systems, and Viktor Kuznetsov for gyro control devices. The preliminary design for the missile called for developing a 170-metric-ton, two-stage missile with a 3000-kg detachable nose cone. The nose cone was designed to hold a nuclear warhead weighing up to 1000 kg and having an 80-kiloton maximum yield.

However, in 1953, the Soviet Union successfully tested a thermonuclear warhead—the “hydrogen bomb,” with a yield a few dozen times that of a nuclear warhead. In late 1953, Korolev was ordered to change the design and increase the throw weight to 3000 kg. This required an increase in the total payload mass to 5500 kg. Major revisions in the missile design were required to maintain the desired range specifications.

The basic changes implied by the increased throw weight requirement were as follows:

1. change launch configuration that used the upper load-bearing members of the side-mounted engines as supporting points for the four launch trusses and removed as the missile begins to climb;
2. switch to a high-thrust, four-chamber propulsion system;
3. use movable steering engines in place of exhaust vanes;
4. increase fuel capacity;
5. increase of 38 metric tons in the thrust of each propulsion system;
6. addition of backup gyro control devices;
7. addition of systems to monitor drainage of tanks and synchronization of fuel consumption by the side-mounted engines;
8. development of a fundamentally new launch system that reduced the load on the missile structure, thereby substantially reducing the weight of the missile.

The Government decree ordering development of the R-7 two-stage ballistic missile was approved on 20 May 1954. A subsequent decree provided a schedule and list of deliverables. Flight testing was scheduled for February 1957. Development and fabrication of all systems and the first few missiles took place during 1955–1956.

The missile design represented a fundamental improvement in structure, frame configuration, dimensions and mass, propulsion-system thrust, control actuators, a new dynamic launch system configuration, and new inertial and radio navigation techniques. The missile itself consisted of four side-mounted, first-stage engine assemblies mounted around the central engine assembly of the second stage. The internal design of the side-mounted engine assemblies and central engine assembly were similar to the single-stage designs then in use. The engines used kerosene and liquid oxygen as propellants. All five engine assemblies operated from the ground up and were started almost simultaneously. The side-mounted engines were turned off upon stage separation, and the second stage remained active.

Each engine assembly was based on a standard four-chamber engine with sea-level thrust greater than or equal to 80 metric tons. This package arrangement required a synchronized tank drainage system and an engine-thrust control system. This missile marked the first use of special steering chambers that moved by electrohydraulic actuators in response to commands issued by the control system.

The development process for this missile included extensive experimental testing of individual systems, in addition to experimental testing of the overall design. Final testing of the control system was performed on special M5RD test missiles based on the R-5 missile. Flight testing of 10 M5RD missiles using the new control-system equipment enabled verification of the apparent-speed-control system, the tank drainage system, and a new telemetry system, including the sensor instrumentation (in particular, a vibrational sensor system).

The R-5R test missiles were used for flight testing telemetry-based missile velocity measurements, using a centimeter-wavelength pulsed radio system, radio-wave attenuation in engine jets, and correctness of the underlying design principles for the radio direction finder. Flight testing three R-5R missiles provided a large amount of experimental material that enabled substantial modifications to the instrumentation fabricated for initial flight/design testing of the R-7 missile.

Almost all of the sensor equipment for the onboard instrument system was developed from scratch. In all, seven sets of recording sensors were installed on board the missile. In addition, special ground-based recording equipment was used to monitor the missile launch and the behavior of launch systems. More than 600 in-flight parameters were measured using a total of 2800 kg of onboard instrumentation. Unlike all previously developed single-stage missiles, the R-7 missile and launch fixtures formed a single dynamic system. The launch ground system included more than 30 individual systems and assemblies.

The interfaces between ground systems and missile and the interfaces between individual ground systems were tested using special missile mockups interfaced to the ground systems at the launch facility. However, before this, appropriate system testing of the launch system and missile was performed at

the Leningrad Machinery Plant. The missile-lift procedure was simulated using high-capacity lifting cranes and a missile filled with water in place of fuel. It was essential to ensure that the gantries retracted simultaneously and that the missile frame would mechanically mate with the launch system. The complete ground system at the launch site was tested using a simulated R-7 missile, which was repeatedly placed on the launch stand and repeatedly fueled.

The Scientific Research Institute for Firing Tests (NII-229) built several high-capacity rocket test stands for testing the propulsion systems in combination with the missile frame. From August 1956 to March 1957, NII-229 performed five firing tests of the side-mounted engine units, three tests of the central engine unit, and two tests of the complete package of five engine units. The first R-7 missile for flight testing arrived at the Tyuratam support facility (Scientific Test Site NIP-5, the future Baikonur) in March 1957. The horizontal testing performed at the NIP-5 support facility included electrical and pneumatic testing of each engine unit, postshipment verification of engine-unit alignment, assembly of the engine units into a single package, and integration testing of all electrical and pneumatic systems ("horizontal system testing").

The first meeting of the State Flight Test Committee occurred on 10 April 1957. The Committee was chaired by M.V. Ryabikov, Chairman of the Military Industrial Complex, and had the following members: Chief Marshal of Artillery M.N. Nedelin (Deputy Chairman); Chief Designer S.P. Korolev (Engineering Manager); Chief Designers V.P. Glushko, N.A. Pilyugin, V.P. Barmin, M.S. Ryazanskii, V.I. Kuznetsov, S.M. Vladimirkii (Deputy Chairman, State Committee for Radio Electronics), A.I. Nesterenko (Manager, Scientific Test Site 5), G.N. Pashkov (State Planning Committee [Gosplan]), I.T. Peresypkin (USSR Ministry of Communications), and G.R. Udarov (Deputy Chairman, State Committee for Military Equipment).

The missile was first launched on 15 May 1957. The flight appeared to proceed normally for the first 60 seconds, at which time a fire broke out in the tail compartment. Reduction of the telemetry data revealed that one of the side-mounted engines fell off 98 seconds into the flight, when the missile became unstable. The root cause of the accident turned out to be a leak in a fuel line. Nevertheless, this launch did confirm that the control-system parameters for the first-stage segment were correct and gave us confidence in the launch dynamics.

The second scheduled launch attempt on 11 June 1957 was unsuccessful because the disc on the main oxygen valve for side-mounted unit C froze and an error had been made during installation of the nitrogen blowdown valve on the oxidant line for the central engine unit. The missile was returned to the Support Facility.

The third launch was on 12 July 1957. Thirty-three seconds into the flight, the missile became unstable. The root cause of the accident turned out to be a short circuit between the control-signal circuits and the housing in the angular-velocity integrator for the roll channel. The fourth launch on 21 August 1957 was successful, and the missile, for the first time, hit a target on Kamchatka Peninsula. A TASS statement announcing the launch of a long-range, multistage intercontinental ballistic missile was carried by the USSR mass media on 27 August.

The fifth launch of the R-7 missile on 7 September 1957 confirmed the results of the previous launch. However, although the missile components and systems operated normally during the active portion of the flight, the warhead

reentry vehicle (RV) broke up upon reentry into the lower atmosphere. A significant amount of time was required for research and development work related to, and fabrication of, new warhead RVs capable of withstanding the high temperatures and large gas-dynamic loads that occurred during the reentry portion of the flight. With the consent of the Government and the State Commission, it was decided to use two of the original twelve R-7 missiles built for development testing as launch vehicles for the first artificial Earth satellites. These launches provided a practical opportunity to accumulate additional experimental data on all missile systems except for the forward compartment containing the nuclear warhead. Thus, these early satellite launches were an integral part and outgrowth of development testing for early intercontinental ballistic missiles.

A fortunate convergence of historical fates in the early 1950s led to the renewal of creative contacts between Sergei Korolev and Mikhail Tikhonravov, his former colleague in development of early perwar unguided missiles during the 1930s. In the late 1940s, Tikhonravov led a group of enthusiasts from the highly classified Scientific Research Institute 4 [NII-4] who were studying designs for space launch vehicles, as well as a variety of issues related to subsequent manned spaceflight. Tikhonravov and his group were the first professional people in the Ministry of Defense system who dared to say that the R-7 intercontinental ballistic missile developed under Korolev's direction could, with slight modifications, be used as a satellite launch vehicle.

With slight changes to the flight plan, the missile could place a satellite weighing up to 1500 kg into orbit instead of delivering a 5.5-metric-ton nuclear warhead at a range of 8000 km. The Ministry of Defense generals did not support Tikhonravov's initiative, but Korolev quickly grasped the possibilities for practical implementation of his long-held dream of human spaceflight, and was able to arrange for Tikhonravov's transfer from NII-4 to OKB-1 (Korolev's design bureau). In May 1954, Korolev had presented a proposal to the Minister of Armaments (Dmitrii Ustinov), the Council of Ministers, and the USSR Academy of Sciences to develop the world's first artificial Earth satellite and launch it with the R-7 missile for research purposes. The basic idea contained in Korolev and Tikhonravov's report was that "the artificial Earth satellite is an inevitable phase in the development of space hardware, following which interplanetary missions will become possible."

In August 1954, the USSR Council of Ministers approved a proposal to study the scientific and engineering issues involved in a spaceflight. Korolev had support from the following senior government officials, as well as the Academy of Sciences: Deputy Chairman of the Council of Ministers V.M. Malyshev, D.F. Ustinov, and Ministers B.L. Vannikov, M.V. Khrunichev, and K.N. Rudnev. Despite the military's fears that development of a spacecraft would be a distraction from the primary tasks involved in developing the R-7 missile, on 30 January 1956, the Council of Ministers approved the development of an unstabilized spacecraft ("Object D") weighing 1100–1400 kg and carrying 200–300 kg of scientific research instrumentation. The Government gave the USSR Academy of Sciences responsibility for developing the scientific research equipment. The required task orders were issued as directives to the Ministry of Defense Technology (the lead ministry) and all other ministries and organizations involved in missile development and production.

OKB-1 and its subcontractor organizations completed their design work in 1956 and then moved on to fabricate Object D, which was to become the first artificial Earth satellite. Object D was to be used for scientific research in the following areas: density and ionic composition of the high-altitude atmosphere, solar particles, magnetic fields, cosmic rays, temperature conditions within the spacecraft itself, braking of the spacecraft in the upper atmosphere, and the accuracy of spacecraft position and orbit determination. The scientific instrumentation and spacecraft onboard systems were to have been powered by solar panels and storage batteries, and the spacecraft would have had an automatic temperature control system. This spacecraft was also to have been the first recipient of an onboard control system with a special programmable timer. In-flight control of the scientific research program was to have been performed via a radio command link from the ground. A special radio telemetry system was to have transmitted the research results to the ground. An extensive network of ground stations, forming a unified command and telemetry system controlled from a single center at NII-4, was to have been constructed.

By late 1956, it was clear that the schedules for fabricating the scientific instrumentation had slipped, and it became uncertain whether the spacecraft could even be launched during the U.N.-sponsored International Geophysical Year. Because press reports indicated that the United States was also preparing to launch a spacecraft early in the International Geophysical Year, Korolev proposed postponing the launch of Object D and launching a very simple spacecraft, the PS, carrying no scientific instrumentation instead.

On 15 February 1957, the Government accepted the proposal by Korolev and the Academy of Sciences to launch an extremely simple unstabilized Earth satellite (Object PS) to verify that the PS could be observed in orbit and that signals transmitted by the PS could be received, as well as to ensure worldwide priority in the space race. The Government would permit the spacecraft launch to occur only after one or two successful launches of the R-7 missile. The extremely simple satellite (the PS) was designed, fabricated, and prepared for launch using the missile in only 8 months. The launch was scheduled for 6 October. However, intelligence reports from overseas indicated that the Americans were preparing their own satellite for launch in early October, so Korolev sped up the preparations, and the PS was launched on 4 October. The first artificial Earth satellite was launched into space on the fifth launch of the first intercontinental missile. Two of the preceding four launches had been unsuccessful because the problems encountered by the developers of the first intercontinental missile were new.

The goal of this launch was to place an extremely simple satellite into orbit and also to obtain additional experimental data on the dynamics of the missile launch and propulsion systems, the control and guidance systems, the stage separation system, the onboard sensors, the operation of equipment on the ground, and the command and telemetry system on the ground. During the round-the-clock effort to prepare the rocket for launch, several problems were identified and eliminated right on the launch pad. In response to one problem report during rocket fueling, one of the fuel tanks for the side-mounted engines was drained and refilled to test the "tank full" alarm system.

The guidance system was adjusted to place the missile into an orbit with the following parameters:

perigee: 223 km;
apogee: 1450 km;
period: 101.5 min.

These orbital parameters could be achieved using half of the guaranteed fuel supply, provided that the guidance systems and all engine systems operated properly. Trajectory measurements indicated that the rocket performed normally during the active portion of the flight. However, the second-stage engine ran out of fuel before it was scheduled to be turned off by the guidance system.

The world's first artificial Earth satellite was launched into space on 4 October 1957 at 22:28 Moscow time. It ended up in an orbit with the following parameters: perigee; 228 Km: apogee; 947 Km: inclination; 65.1° : and period: 96.17 min. Reduction of the telemetry data enabled collecting a large amount of experimental data concerning operation of the missile, various individual missile systems, and the ground launch systems. The spacecraft weighed 83.6 kg; the body of the spacecraft consisted of a sphere 580 mm in diameter with four collapsible-whip antennas (2.4 m and 2.9 m long). The body of the first artificial Earth satellite consisted of two aluminum alloy hemispheres filled with dry nitrogen (pressure 0.13 MPa); the joint between the two hemispheres was sealed using a rectangular-cross-section, vacuum-grade O-ring. The pressurized enclosure held an electrochemical power source and two radio transmitters that continuously transmitted on 20.005 and 40.002 MHz (wavelength 17 and 7.5 m, respectively). The transmitter signals consisted of alternating telegraph "marks" and "spaces," each 0.3 s long. The "mark" on each frequency was transmitted simultaneously with the "space" on the other frequency. The radio transmitter system had a total mass of 3.5 kg, and each transmitter had an output of about 1 W. The telemetry data (temperature and interior pressure) were transmitted to the ground by modulating the frequency of the "mark" and "space" signals. Each transmitter had two collapsible-whip antennas (approximately 70° apart). Each pair of antennas had a nearly spherical antenna pattern.

The temperature control system had a radiator with a fan-driven, sealed-loop, forced-gas heat exchange system designed to maintain a stable interior temperature in the face of variable external thermal fluxes. The temperature control system used a bimetallic thermal relay as the sensor element. Whenever the temperature increased above 36°C , the fan came on, and nitrogen circulated through the system to transfer heat away from the hemisphere that was acting as a radiating surface (emission coefficient 0.35–0.4, solar absorption coefficient 0.23–0.4). The fan was turned off whenever the temperature fell below 20°C .

The intended purpose of the automated onboard electrical system was to turn on the electrical power to the instruments, once the spacecraft reached orbit (i.e., upon separation from the launch vehicle). During the launch phase, the spacecraft was placed under a nose cone for protection against aerodynamic and thermal effects; the nose cone was jettisoned when the second-stage engine shut down. The spacecraft carried three silver-zinc batteries weighing 51 kg. The batteries could support operation of the instruments for 3 weeks. The first

satellite lasted 92 days, and completed ~ 1400 orbits around Earth. On 4 January 1958, it reentered Earth's atmosphere and burned up.

The orbital parameters of the first Soviet artificial Earth satellite were such that it was visible from all continents across a wide range of latitudes. Observations of the spacecraft's motion, reduction of the observations, and prediction of the future motion of the spacecraft based on these results served as an early practical exercise in using ground-based spacecraft control systems and equipment for measuring spacecraft parameters. The first satellite was observed using radio equipment, as well as optical instruments at astronomical observatories. The news that the first artificial Earth satellite had been launched aroused extremely broad interest among radio amateurs and amateur astronomers around the world. In the Soviet Union, 66 optical observing stations and 26 clubs with extensive collections of radio observing gear regularly observed the spacecraft. In addition, thousands of radio amateurs attempted radio observations of the spacecraft. State scientific radio monitoring stations also observed with radar and radio range finder systems. This was the first rich set of statistical data concerning the transmission of meter-wavelength radio waves through the ionosphere and also represented the first opportunity to receive radio signals at two different frequencies from regions of the ionosphere that had heretofore been inaccessible, that is, above the ionization peak or even above the entire ionosphere.

Extremely valuable data were also obtained on radio-wave absorption in previously unstudied layers of the ionosphere, as well as new data on the structure of these regions and the ion concentration at various altitudes and times of day. These systematic measurements showed that the altitude of the main peak in the ionosphere and the peak electron concentration vary from day to night, north to south, and east to west. Radio propagation measurements at the frequencies emitted by the spacecraft at various altitudes provided a new avenue for ionospheric research. These observations led to the discovery that the decrease in electron concentration in the upper ionosphere (above the main peak) with altitude is five to six times slower than the increase with altitude below the peak. For example, when the observations were made (in October), the electron concentration increased by approximately a factor of 10 from 100–300 km altitude, whereas it decreased only by a factor of 2 from 300–500 km in altitude. The initial indications that micrometeoroids were not hazardous to spacecraft was an extremely important result.

The radio methods used included radio ranging and Doppler observations of the radio signal from the spacecraft. These early experiments indicated that the Doppler effect could be used successfully to determine spacecraft orbital parameters. It became obvious that the accuracy of orbit determination could be quite high if the transmitter frequency were increased and if automated frequency measurement equipment were used. Early, high-sensitivity photographic techniques were developed for observing this spacecraft. Image-converter tubes turned out to be especially promising in this regard.

The news that the Soviet Union had launched the world's first artificial Earth satellite turned out to be quite an unexpected sensation for the entire human community. The flight of the first satellite around Earth caused a stunning resonance around the world. Virtually the entire world press carried front-page banner headlines reporting the news, thereby underlining that Soviet

science had taken the lead. The American government was shocked, along with American scientists confident of their superiority. The senior leadership of the Soviet Union was surprised by this enthusiastic reaction from their people and people around the world. Therefore, to consolidate the political success that had been so unexpectedly achieved in the Cold War, the General Secretary of the CPSU Central Committee, Nikita Khrushchev, proposed that Korolev launch a new satellite in honor of the 43rd anniversary of the October Revolution. The first, extremely simple satellite was still operating in orbit, and there was no sense in launching another, similar satellite.

A second spacecraft was readied in less than a month and launched on 3 November 1957. The dog Laika, who later became famous, was the first experimental animal to orbit Earth. The second spacecraft did not separate from the second stage of the rocket, and, for the first time, data on the behavior of a dog in space was transmitted to the ground via a multichannel telemetry link.

Object D, which was to have been the first spacecraft, was not launched until 15 May 1958. It was the third Soviet spacecraft and the first that could truly be called a space laboratory based on the amount of scientific instrumentation on board.

The successful launch of the first artificial Earth satellite marked the beginning of humanity's journey into space. Many extremely urgent scientific problems required direct experiments at altitudes of hundreds or thousands of kilometers above Earth's surface. Although the significance of artificial Earth satellites had long been understood, launching them had remained an insoluble problem. The main difficulty had been developing a rocket that could give a spacecraft a velocity of the order of 8000 m/s. Only after the Soviet Union had developed an intercontinental ballistic missile did it become possible for the first time in history to launch an artificial Earth satellite. The superior design of this missile enabled placing a spacecraft with the required weight of scientific instrumentation into orbit. Further improvements of the R-7 missile—in particular, the addition of a third and then a fourth stage led to manned space programs, communications satellites, and the first automated interplanetary spacecraft for studying the Moon, Venus, and Mars. Modernized versions of the R-7 missile are currently used for manned and cargo spacecraft to support the International Space Station.

BORIS CHERTOK
ENERGIA Space Association
Russia

SUN

THE SUN AS A STAR

Although many cultures deified the Sun, for example, Ra, the Egyptian sun god, and SŪRYA, the sun god in Ancient India, humans became aware of its real

significance to terrestrial beings only in the last 400 years. Before the Copernican heliocentric model of the solar system was accepted, the Sun and the “wandering stars” (which we now know are planets), the Moon and the fixed stars were objects whose apparent daily motions needed to be explained. The Ptolemaic system hypothesized intricate moving equants and epicycles to describe the motion. By placing the Sun logically at the center of the solar system and observing moving spots on the Sun, Galileo established its place in our current world view. Contemporary humans have learned that the Sun supplies a nearly constant flux of visible radiant energy to Earth and is also magnetically connected to it so that powerful releases of energy at a distance of 1 AU (1.5×10^8 km or ~ 93 million miles) from Earth can produce surprisingly large local effects, such as the extensive power failure in October 1989.

Basic Properties of the Sun. The Sun is only one of several billion stars in our galaxy, but as a special example, it is important to understand how it compares to other stars. Following the development of stellar spectroscopy, techniques became available to classify stars by their locations on the Hertzsprung–Russell (HR) diagram, which is a plot of stellar luminosity as a function of spectral type. Figure 1 from Chaisson and McMillan (1) shows an HR diagram for several varieties of stars. The ordinate is the object’s optical luminosity or the total energy emitted per second by the object, and the abscissa is the star’s spectral type, which is a measure of its surface temperature. To place a star on the HR diagram, the star’s distance must be known, so its absolute luminosity can be determined from the measured luminous flux using the inverse square law. As in Fig. 1, the luminosity is usually expressed in terms of the Sun’s luminosity. The star’s spectral type is determined by measuring the star’s “color” and characterizing its spectrum. The sun is a “G” class star. Stars that have higher surface temperatures, for example, are “O” class, and “M” class stars have lower surface temperatures. All normal stars span a temperature range from $\sim 40,000$ K to ~ 4000 K. Actually a star’s spectral class is determined by comparing its spectrum with the spectra of a set of standard stars. Some physical properties of the Sun and the way the properties are determined are shown in Table 1 (from Reference 2).

Of most interest here is the group of stars on the main sequence, the so-called “normal stars,” that are in the early stages of their evolution. (The Sun is indicated by the symbol \odot .) Formally, the Sun’s spectral class is G 2 V, and its surface temperature is ~ 6000 K. It is considered an average main sequence star in all respects: luminosity, surface temperature, and size. The M class stars are typically 0.1 times smaller than the Sun, and the O class stars are 100 times larger than the sun. However, due to the sun’s proximity to us, detailed studies of it and the radiations it emits can be done with great precision, and this knowledge can be used to understand other stars better.

From direct measurement of isotope ratios, we know that the Sun’s age is ~ 4.5 billion years, and from the theory of stellar evolution, that it may live at least another 5 billion years. When its current internal energy source is depleted, it will expand and become a red giant and eventually a dwarf star similar to those on the HR diagram in Fig. 1. As far as its origin is concerned, the Sun and planets may have formed as a result of a supernova explosion near the early solar nebula. This aspect of the Sun’s early life and its future are fascinating topics

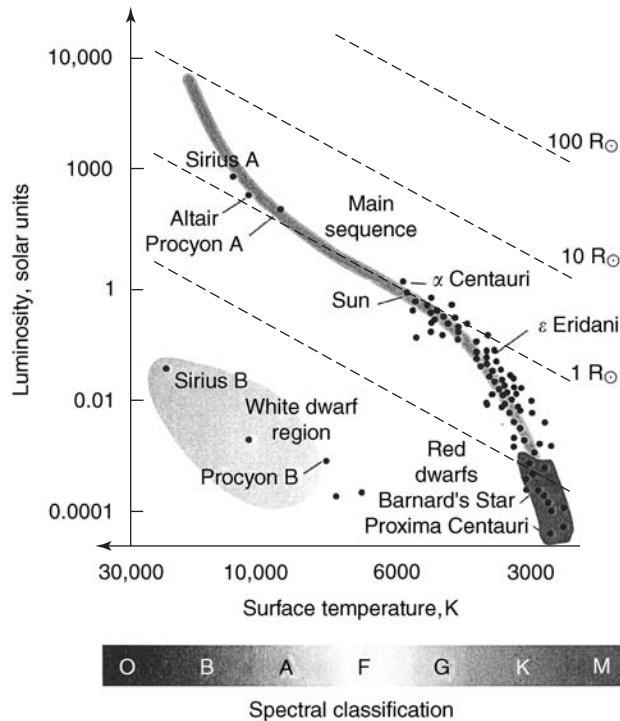


Figure 1. Most stars have properties within the shaded region known as the main sequence. The points plotted here are for stars lying within about 5 pc of the Sun. The diagonal lines correspond to constant stellar radius, so that stellar size can be represented on the same diagram as luminosity and temperature. (Recall that \odot stands for the Sun.) From Reference 1.

(see, for example, Reference 3), but our goal here is to describe the present Sun in some detail.

The current understanding of the Sun draws on advances in both ground-based observing techniques and sophisticated space instrumentation, as well as in theoretical computer models using the observations as input. This review first briefly describes some important ground-based observations and identifies several space missions whose main purpose has been to advance our knowledge of our nearest star. Next, the structure of the Sun and its atmosphere is described in more detail, beginning at the center of the Sun and emphasizing our current understanding in appropriate detail. A major goal of this work is to give an up-to-date summary of our knowledge of solar activity. The new science of helioseismology is discussed later, and space weather, as it affects Earth, is also covered briefly later. For each topic, a brief history will be given with a summary of current knowledge followed by identification of important unsolved problems. References for further study will be given for each topic discussed. A general reference for the topics discussed can be found in chapter 16 of *Astronomy Today* by Chaisson and McMillan (1).

Table 1. **Properties of the Sun**^a

Datum	How found	Value
Mean distance	Radar reflection from planets	1 AU
Maximum distance to Earth		149,597,892 km
		1.521×10^8 km
Minimum distance to Earth		1.471×10^8 km
Mass	Acceleration of Earth	333,400 Earth masses
Mean angular diameter	Direct measure	1.99×10^{33} g 31'59".3
Diameter of photosphere	Angular size and distance	109.3 times Earth diameter
Mean density	Mass/volume	1.39×10^{11} cm 1.41 g/cm^3
Gravitational acceleration at photosphere (surface gravity)	GM/R ²	27.9 times Earth surface gravity
Solar constant	Measure with instrument such as bolometer	27,300 cm/s ²
Luminosity		1.9 cal/min/cm^2
Spectral class	Solar constant times area of spherical surface 1 AU in radius	$21.368 \times 10^6 \text{ ergs/s/cm}^2$ $3.8 \times 10^{33} \text{ ergs/s}$
Effective temperature	Spectrum	G2V
Visual magnitude	Derived from luminosity and radius of Sun	5800 K
Apparent	Received visible flux	− 26.7
Absolute	Apparent magnitude and distance	+ 4.8
Rotational period at equator	Sunspots and Doppler shift in limb spectra	24d16h
Inclination of equator to ecliptic	Motions of sunspots	7°10'.5

^aFrom Reference 2.

Ground-Based Observations of the Sun. Heracleides (388–315 B.C.) and Hipparchus (190–120 B.C.)¹ can be considered the first who observed the Sun with a scientific motivation. The Ptolemaic Earth-centered system, as promoted by Aristotle, held sway until Copernicus (1475–1543) proposed the heliocentric model for the solar system, even though Aristarchus (~320 ~ 250 B.C.) had earlier considered such a model. However, serious study of the Sun began after Lippersheim invented the telescope in 1608. Galileo (1610) in Padua and Fabricius (1610) in Wittenberg used it in their independent discoveries of sunspots. Solar physics investigations for the next 192 years consisted of recording the number of sunspots on the solar disc until 1802 when Wollaston discovered dark lines in the solar spectrum. Using an improved spectrograph, Fraunhofer discovered 547 additional dark lines, features that were named after him. The origin of these strange features on the Sun, as explained in 1859 by Kirchhoff and Bunsen, is the selective absorption of light from the hot solar surface that is directed toward Earth, by the cooler atoms in the Sun's atmosphere. By 1929, the chemical composition of the Sun was known as a result of the pioneering work of Hale and Russell and many others. That the Sun rotates in a period of about a month was first inferred from the observation of sunspots. A more complete summary of the knowledge of the Sun's properties, obtained before the era of space age astronomy, can be found in several readable texts such as Menzel (4) and Kiepenheuer (5). *Early Solar Physics* by Meadows (6) contains an interesting review of ideas about the Sun in the period 1850–1900 and has several original papers. Hufbauer (7) traces the development of knowledge about the Sun from Galileo's time.

Space-based observations of the Sun are complemented by very significant ground-based observations. From the latter group, we discuss the following:

- The discovery of radio emissions from the Sun in 1942 during the development of radar, which led to the study of solar radio bursts by using radio-heliographs.
- The discovery of neutrinos from the Sun whose flux was significantly below theoretical expectations; this discrepancy is driving a revision of our understanding of stellar interiors and also fundamental particle physics. [See **notes added in proof.**]
- The development of helioseismology, beginning with the discovery of 5-minute oscillations of the Sun in 1960, which led to the development of the Global Oscillation Network Group (GONG) project, the Michelson Doppler Imager/Solar Oscillation Investigation (MDI/SOI), and, the Extreme-Ultraviolet Imaging Telescope (EIT) instruments on the Solar and Heliospheric Observer (SOHO) satellite.

Observations of the Sun from Space. Table 2 is a chronological list of the principal space missions, which have, or are expected to, provide significant advances in our knowledge of the Sun. Here we give a brief summary of the main facets of the Sun that a given mission is intended to study. In the following

¹See Asimov's *Biographical Encyclopedia of Sciences and Technology*, Doubleday, Garden City, 1972, for biographical sketches of ancient to modern scientists.

Table 2. List of Space Missions Relevant to Solar Physics^a

Mission name	Launch date	Termination date
Sounding rockets	1946	Continuing
Vanguards	1959	
Solrads	20 June 1960	1975
Orbiting Solar Observatories (OSO)	1962	1975
Mariner 2	27 August 1962	14 December 1962
Skylab	1973	1974
Helios 1	1974	1984
Helios 2	1976	1981
ISEE3/ICE	1978	1982
P78-1	1978	1985
Solar Maximum Mission (SMM)	14 February 1980	3 December 1989
Hinotori	21 February 1981	
Space Shuttle	1982	Continuing
GRANAT	1 December 1989	1974
Ulysses	6 October 1990	Continuing
Gamma-1	1991	
Compton Gamma-Ray Observatory (CGRO)	5 April 1991	4 June 2000
YOHKOH	30 October 1991	Continuing
Solar and Heliospheric Observer (SOHO)	2 December1995	Continuing
Advanced Composition Explorer (ACE)	25 August 1997	Continuing
Trace	2 April 1998	Continuing

^aThe National Oceanic and Atmospheric Administration (NOAA) maintains regular observations of solar emissions, X-rays, and particles, with its GOES satellites. [See notes added in proof.]

sections, we describe in more detail some of the observations or discoveries made on these missions.

The first observations of the Sun from space were made by Naval Research Laboratory (NRL) scientists in 1946 using ultraviolet (UV) and X-ray spectrometers carried on V-2 rockets captured from the German military in 1945. Further observations were made of solar UV and X-ray emissions using Viking and Aerobee rockets. After 1980, Black Brant rockets were used as well as the Aries and the second stage of the Minuteman I ICBM. Sounding rockets of various types are still used today to obtain—mainly UV, extreme UV (EUV), and X-ray spectra of the Sun using precise imaging telescopes. Sometimes, sounding rocket instruments are used for cross-calibration of instruments carried on satellites such as SkyLab or the Solar Maximum Mission. (A discussion of some of the early results is given in Reference 8.)

The first satellite observations of solar Lyman alpha (1216 Å or 121.6 nm) and soft X rays (1–8 Å or 10–80 nm) were made by NRL from SolRad I during a period of 5 months after launch on 10 June 1960. Additional SolRad satellites produced continuous data on UV and soft X rays through 1975.

Starting in 1962, NASA launched its Orbiting Solar Observatory (OSO) Series, which made many important observations of the quiet and active Sun through 1975. The first white light coronagraph and high-resolution gamma-ray spectrometer were flown on OSO 7, launched in 1971 and operated for 15 months.

During its 4-month journey to Venus in 1962, Mariner 2 made comprehensive and definitive observations of the solar wind.

Following OSO 7, the SkyLab (the last phase of the manned Apollo program) was launched in 1973 and continued into 1974. Imaging soft X-ray and EUV telescopes made pioneering observations related to the solar wind (SW) and coronal holes. (See also Reference 9 for a brief review of early results from the SolRads, OSOs, and SkyLab.)

From 1974 to 1976, the European Helios 1 and 2 solar-orbiting satellites made plasma and high-energy electron and ion observations at distances from the Sun as close as 0.3 AU.

From 1978 to 1982, the International Sun Earth Explorer 3 (ISEE3) was at the Sun–Earth Lagrangian point (L1) and made numerous valuable solar flare observations. As the International Comet Explorer (ICE), it continued flare observations beyond 1982.

In 1978, the U.S. Air Force launched the P78-1 satellite that carried a white light coronagraph (Solwind) and observed many coronal mass ejections (CMEs). In the early 1980s, the P78-1 mission unceremoniously ended when it became a target in a “Star Wars” test.

On 14 February 1980, the Solar Maximum Mission (SMM) satellite was launched. It carried six instruments for a coordinated study of solar flares, at the maximum of sunspot cycle 21, covering several spectral regions: the UV, soft and hard X rays with the hard X-ray burst spectrometer (HXRBS), and gamma rays, with the gamma-ray spectrometer (GRS). The instrument payload included a cavity radiometer for measuring the total solar luminosity or “solar constant.” In April 1984, SMM was taken onboard the space shuttle Challenger to replace gyroscopes and to refurbish some instruments. SMM was returned to orbit and made important flare observations during the rising phase of sunspot cycle 22. Its mission was ended on 2 December 1989 by orbital decay caused by expansion of earth’s atmosphere due to heating from solar UV and X-ray emissions, the very radiations it was designed to study! Early results from all SMM instruments are described in a report of an SMM workshop (10).

The Japanese Hinotori solar-observing satellite, launched on 21 February 1981, made many observations of solar flares, some of which overlapped SMM/GRS observations. (A review of high-energy solar flare observations during this period is given in Reference 11.)

In 1982, the manned Space Shuttle program began. Several small special-purpose solar observations were made from the SPARCS satellites, which were developed earlier for sounding rocket flights. SPARCS was deployed during a 7 to 10-day mission by shuttle astronauts, coorbited with the shuttle, and then was retrieved at the end of mission. It was returned to Earth, and a new solar instrument was installed, ready for another shuttle launch. The major Shuttle mission of importance to solar physics was the launch of Spacelab 2 on 29 July 1985. Spacelab 2 carried three solar telescopes and several other instruments with two solar physicists, Loren Acton and John-David Bartoe, to operate the solar telescopes. (See Reference 12 for a complete history of the Spacelab program.)

The Russian Granat satellite was launched in December 1989 carrying the French/Russian SIGMA, a new type of imaging gamma-ray telescope and a

scintillator gamma-ray spectrometer, PHEBUS, that responded omnidirectionally. Therefore, PHEBUS could detect many gamma-ray flares, but not with the same efficiency as the SMM/GRS.

In October 1990, the ESA Ulysses was launched. Its mission is to study the high-latitude heliosphere, that is, above and below the plane of the ecliptic. It also carried instruments to study the solar wind composition and charge state and an X-ray spectrometer to study solar flares. A review of the results from the first orbit of Ulysses around the Sun was presented by Reference 13.

The Russian spacecraft GAMMA-1, operational during 1991, made valuable observations of two high-energy flares.

On 5 April 1991, the Compton Gamma-Ray Observatory (CGRO) was launched. Its primary mission was to study cosmic sources of gamma rays with four instruments, which covered the photon energy range from a few keV to 30 GeV. However, as a secondary goal, all CGRO instruments can make solar flare observations and have made some significant discoveries. On 3 June 2000, CGRO underwent a controlled reentry over the Pacific Ocean to reduce the chance of debris impacting populated areas.

The Japanese YOHKOH solar-observing satellite, launched on 31 August 1991, carried the soft and hard X ray imaging telescopes (SXT) and (HXT). The dramatic imaging capabilities of the SXT and HXT have revolutionized the study of solar activity.

On 2 December 1995 the Solar and Heliospheric Observer (SOHO) was launched with 12 instruments for studying several phenomena of the quiet and active sun. Helioseismology, solar wind (SW) and coronal mass ejection (CME) observations from SOHO are discussed in individual sections later.

The Advanced Composition Explorer (ACE), launched on 25 August 1997, is designed to study solar flare particle emissions, cosmic rays, the SW, and other heliospheric energetic particle phenomena. It is positioned between the Sun and Earth at the Sun/Earth neutral gravity point, L1. (For solar flare particle emissions, see the section on Space Weather, in this article.)

The Transition Region and Coronal Explorer (TRACE) was launched on 2 April 1998 to image (to 1 arcsecond spatial resolution) the photosphere, the transitional region, and the corona at three EUV wavelengths and several UV wavelengths. The solar plasma can be studied at selected temperatures from 6000 K to 10^7 K.

STRUCTURE OF THE SUN

The structure of the Sun is illustrated in Fig. 2 as a series of concentric spherical shells, or zones, which are understood to have different physical characteristics, as deduced from observations and the theoretical development of the standard solar model (e.g., the SSM described by Bahcall and Ulrich in (14a) and later by Bahcall and Pinsonneault (14b). The inner zone or core of radius $\sim 2 \times 10^5$ km, the region of energy generation by nuclear burning, is followed by the radiative zone of outer radius $\sim 5 \times 10^5$ km and then the convective zone of outer radius $\sim 7 \times 10^5$ km. The thin outer zone of the solar disk, which is called the photosphere, has an accurately determined thickness of $\sim 5 \times 10^2$ km and a

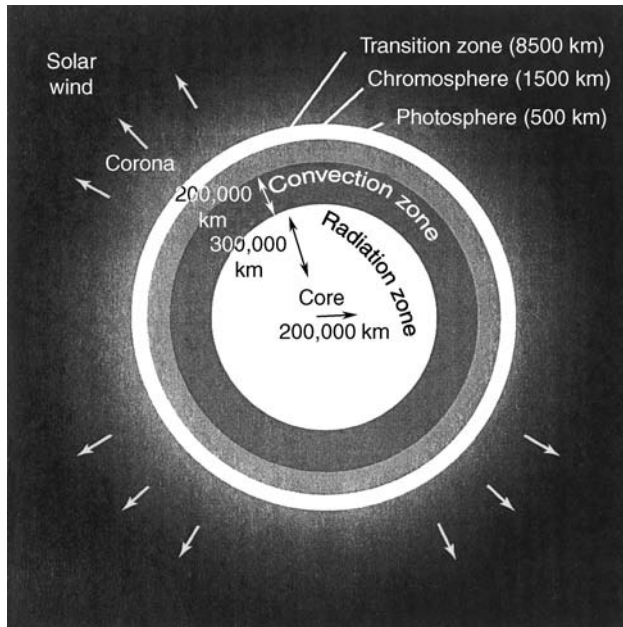


Figure 2. A schematic of the Sun showing the relative location of various regions and their approximate thicknesses. Not drawn to scale. From Reference 1.

temperature of ~ 5800 K. These zones lie within the Sun's visible disk as measured at a wavelength of 500 nm, although we see only the photosphere. Figure 3, from Reference 15, shows the internal physical properties of the Sun predicted by the SSM. A new means of studying the interior of the Sun uses helioseismology techniques on the SOHO satellite that complement the GONG observations.

Above the photosphere lies the solar atmosphere that has three distinct regions: the "cool" chromosphere (~ 4500 K) $\sim 1.5 \times 10^3$ km thick, sometimes called the color sphere and seen in some total eclipses followed by the transitional region ~ 8500 km thick in which the temperature rises rapidly to ~ 8000 K, and then the low-density, hot corona at a temperature of $\sim 1 \times 10^6$ K, also seen in a total eclipse. The expansion of the corona into the interplanetary medium (IPM) becomes the solar wind, which extends throughout the solar system.

The Core and the Neutrino Problem. It was realized in the mid-1800s that if the source of the Sun's energy was due to its gravitational contraction, the so called "Helmholz–Kelvin hypothesis," then the Sun could not have existed longer than $\sim 3 \times 10^7$ years. However, by the 1930s, age dating of terrestrial rocks containing radioactive minerals indicated that they had existed for $\sim 1.6 \times 10^9$ years., so clearly gravitational energy is inadequate to account for the existence of the Sun and, in fact, all stars on the HR main sequence. That transmutation of elements might be a source of stellar energy was considered as early as 1920 by Eddington and others (see Reference 16). Numerous authors investigated the energy release from possible nuclear reactions, but a paper by Bethe in 1939 (17) provided details for the fusion, or nuclear burning, of

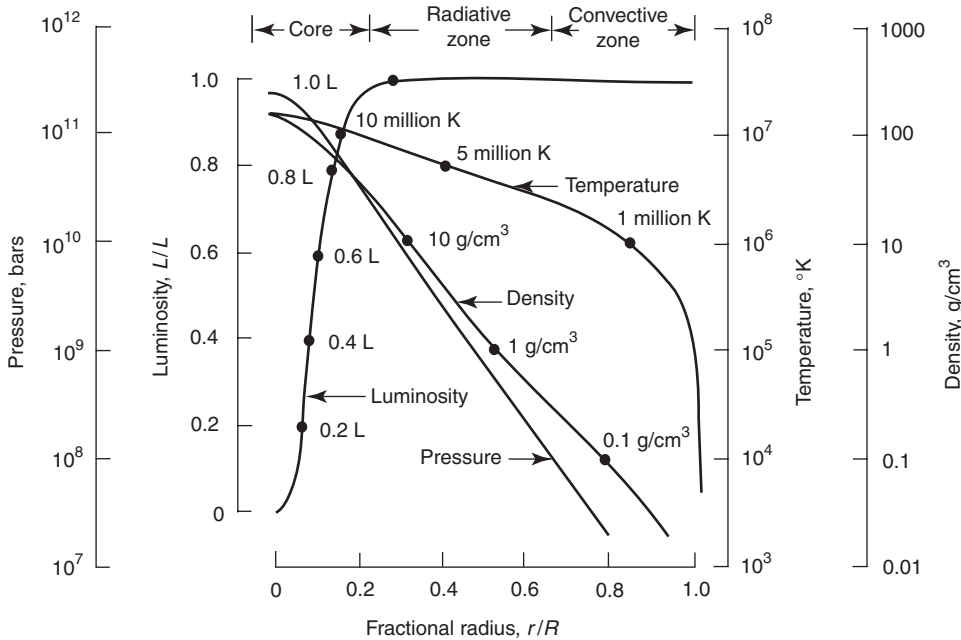


Figure 3. Internal compression. The variation of pressure, luminosity, temperature, and density with fractional radial distance from the Sun's center (left) to its visible surface (right). At the Sun's center, the temperature is 15.6 million Kelvin, and the density is 151 grams per cubic centimeter; the central pressure is 233 billion times that of the Earth's atmosphere at sea level (one bar). Thermonuclear energy is produced in a core region that extends to about one-quarter (0.25) of the solar radius; the core contains almost half of the Sun's mass. The convective zone begins at 0.71 of the solar radius where the temperature has dropped to about 2 million Kelvin and the density has fallen to about 0.2 grams per cubic centimeter; the convective zone comprises about 2% of the Sun's mass. In the photosphere, the temperature is 5780 Kelvin, and the pressure and density have dropped off the scales of the graph. [Prepared from the standard solar model data computed by John Bahcall and Marc H. Pinsonneault, *Rev. Mod. Phys.* 64: 885–926 (1992). From Reference 15.]

hydrogen to helium via the CNO cycle and the p-p chain. These two pathways for producing the Sun's energy are shown in Table 3, although there are three alternative p-p chains. The basic CNO cycle, in which the heavy elements act only as catalysts, was suggested in 1938 independently by Bethe and von Weizsaecker. These reactions convert four hydrogen atoms into one helium nucleus that results in a mass loss of $\sim 4.8 \times 10^{-29}$ kg, or an energy release of 4.3×10^{-12} J, or 26.8 MeV, for every helium nucleus produced by "burning" $\sim 6.7 \times 10^{-27}$ kg of hydrogen. From this basic fact one can estimate that the solar luminosity of $\sim 4 \times 10^{-26}$ J/s would require consuming $\sim 6.2 \times 10^{-11}$ kg of hydrogen per second. Considering the Sun's mass of $\sim 2 \times 10^{-30}$ kg, which is currently mostly hydrogen, nuclear burning could easily power the sun for billions of years. In a sense, the transmutation of hydrogen into helium is very inefficient because only about 0.7% of the available mass can be converted into energy to supply the solar luminosity; however, no more efficient energy source

Table 3. The pp Chain and the Carbon–Nitrogen Cycle^a

Reaction number	Reaction	Neutrino energy, MeV
Proton–Proton Chain		
I-1	$\text{p} + \text{p} \rightarrow {}^2\text{H} + \text{e}^+ + \nu_{\text{e}}$	0–0.420
I-2	$\text{p} + \text{e}^- + p \rightarrow {}^2\text{H} + \nu_{\text{e}}$	1.44
I-3	${}^2\text{H} + \text{p} \rightarrow {}^3\text{He} + \gamma$	
I-4	${}^3\text{He} + {}^3\text{He} \rightarrow {}^4\text{He} + 2\text{p}$	
II-5	${}^3\text{He} + {}^4\text{He} \rightarrow {}^7\text{Be} + \gamma$	
II-6	$\text{e}^- + {}^7\text{Be} \rightarrow {}^7\text{Li} + \nu_{\text{e}}$	0.816, 0.383
II-7	${}^7\text{Li} + \text{p} \rightarrow {}^4\text{He} + {}^4\text{He}$	
III-8	$\text{p} + {}^7\text{Be} \rightarrow {}^8\text{B} + \gamma$	
III-9	${}^8\text{B} \rightarrow {}^8\text{Be} + \text{e}^+ + \nu_{\text{e}}$	0–14
	${}^8\text{Be} \rightarrow {}^4\text{He} + {}^4\text{He} + \gamma$	
Carbon–nitrogen cycle		
	$\text{p} + {}^{12}\text{C} \rightarrow {}^{13}\text{N} + \gamma$	
	${}^{13}\text{N} \rightarrow {}^{13}\text{C} + \text{e}^+ + \nu_{\text{e}}$	0–1.20
	$\text{p} + {}^{13}\text{C} \rightarrow {}^{14}\text{N} + \gamma$	
	$\text{p} + {}^{14}\text{N} \rightarrow {}^{15}\text{O} + \gamma$	
	${}^{15}\text{O} \rightarrow {}^{15}\text{N} + \text{e}^+ + \nu_{\text{e}}$	0–1.73
	$\text{p} + {}^{15}\text{N} \rightarrow {}^{12}\text{C} + {}^4\text{He}$	

^aFrom Reference 23.

has been suggested. A direct experimental proof that the nuclear burning is currently in progress seems to have been presented now (**see later and note added in proof**).

Because the weak interactions in the p–p chain produce neutrinos, it was clear (see e.g., Reference 18 and 23) that detecting them with a flux consistent with theoretical predictions could confirm that nuclear fusion reactions were the source of the Sun’s energy. All of the reaction cycles shown in Table 3 produce neutrinos, but these were not all considered in the early papers about nuclear fusion as the source of stellar energy (see, e.g., Reference 17). The net result of hydrogen burning is $4 {}^1\text{H} \rightarrow {}^4\text{He} + 2e^+ + 2\nu_e$. Because the theory of beta decay was untested (the neutrino was not observed until 1956 at the Oak Ridge nuclear reactor), Pontecorvo (in 1946) and Alvarez (in 1949) proposed detecting neutrinos by the reaction ${}^{37}\text{Cl}(\nu_e, e^-){}^{37}\text{Ar}$. However, the possibility that (the expected) solar neutrinos could be detected eventually led Davis to develop a large neutrino detector consisting of $\sim 100,000$ gallons of cleaning fluid (CCl_4).²

A history of the development of this neutrino detector, other experimental possibilities, and related theoretical work is described well by Bahcall and Davis (19). To understand how this detector relates to the neutrinos produced by nuclear fusion, according to the Standard Solar Model (SSM), Bahcall and Shaviv (20) had deduced the neutrino flux at Earth versus the energy of the neutrinos

²Note that the neutrinos detected by chlorine are “electron neutrinos,” ν_e , but there are at least two other types of neutrinos (see below).

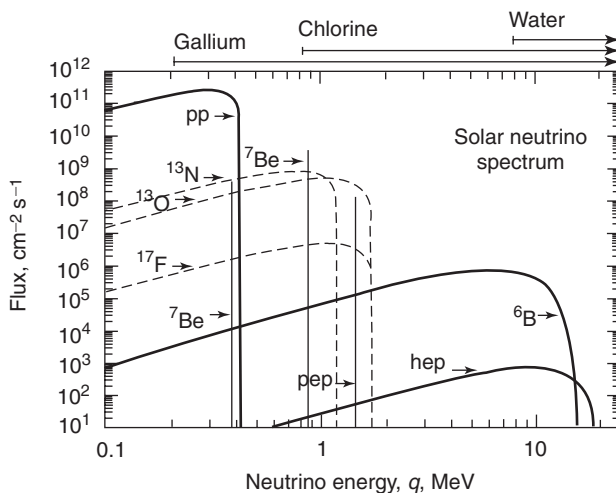


Figure 4. Solar neutrino spectrum. This figure shows the energy spectrum of neutrinos predicted by the standard solar model. The neutrino fluxes from continuum sources (like pp and ${}^8\text{B}$) are given in the units of number per cm^2 per second per MeV at one astronomical unit. The line fluxes (pep and ${}^7\text{Be}$) are given in number per cm^2 per second. The spectra from the pp chain are drawn with solid lines; the CNO spectra are drawn with dotted lines. The upper abscissa scale shows the energy range, from threshold, to which three types of neutrino detectors are sensitive. From Reference 21.

from the different reactions shown in Table 3 before the first results from the chlorine detector had been reported. As can be seen in Fig. 4, from Bahcall (21), the chlorine detector can respond to all neutrinos whose energies are above 0.814 MeV, which can come from ${}^8\text{B}$, from electron capture by ${}^7\text{Be}$, or from the rare proton–electron–proton (pep) reaction.

The chlorine detector has been operated, from 1967 to 1993, in the Homestake gold mine in South Dakota. The first results from the chlorine detector (22) showed that the CNO cycle produces less than 9% of the Sun’s energy. Later results are shown in Fig. 5, a plot of the measured flux of neutrinos from late 1986 to late 1992 (23). The average value of the measured neutrino flux was 2.32 ± 0.22 SNU, where the solar neutrino unit (SNU) corresponds to one neutrino capture per second per 10^{36} atoms of ${}^{37}\text{Cl}$ ³. This flux corresponds to a production rate of 0.437 ± 0.042 atoms of ${}^{37}\text{Ar}$ per day, the units on Fig. 5. The most recent theoretical neutrino flux value, based on the current solar model, is $9.3^{+1.4}_{-1.2}$ SNU. So there is a clear discrepancy, raising the question, Is the experiment or the SSM and basic neutrino physics wrong?

This question and results from three new neutrino experiments, in addition to the Davis results, are discussed in Bahcall (21). Because all of the experiments disagree with the theory, three solar neutrino problems are identified. The discrepancy between the chlorine result and theory is the “first” solar neutrino problem. Now, as shown in Fig. 4, neutrinos whose energy is higher than 7.5 MeV

³To consider detection of neutrinos by target nuclei other than ${}^{37}\text{Cl}$, 1 SNU is more generally defined as 10^{-36} interactions per target atom per second ((21), p. 201).

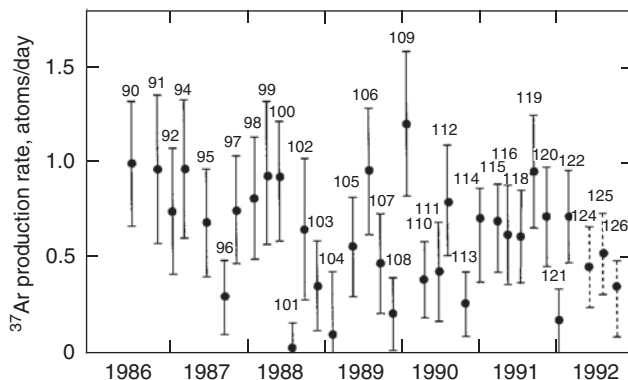


Figure 5. A plot of the ^{37}Ar production rates versus time. This plot summarizes the data obtained after the Homestake experiment resumed operation in September 1986. Run numbers are shown above each plotted point. Not plotted are run 93 for SN1987A and run 117 for the solar flare of 4 June 1991. From Reference 23.

from ^8B decay can be detected by neutrino–electron scattering. Such a reaction can be detected by the Kamiokande II (K II) 4500-ton pure water detector, that records the (Cherenkov) light in a large array of photomultipliers from electrons knocked out of H_2O molecules by neutrinos. After 1000 days of operation, this experiment reported in 1991 a neutrino flux of 0.44 ± 0.06 SNU, about 2.5 times below the predicted value of $1.0^{+0.17}_{-0.14}$ SNU. So both experiments confirm a deficit in the expected solar neutrino flux. However, the chlorine and Kamiokande results are incompatible with one another, and Bahcall identifies this as the “second” solar neutrino problem (see the left and central bar graphs in Fig. 6).

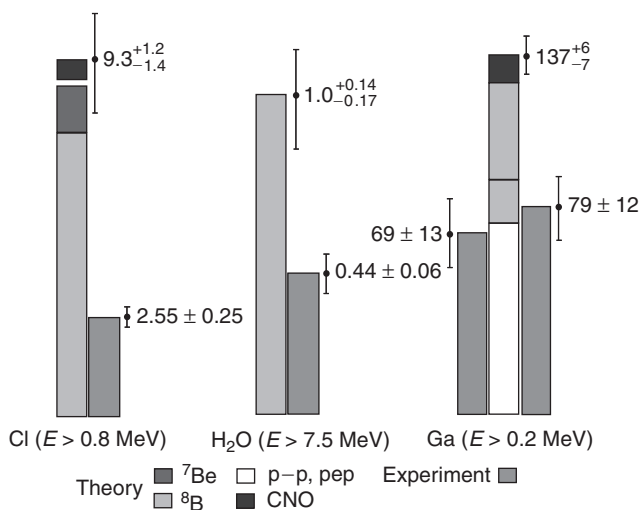


Figure 6. Comparison of measured rates and standard model predictions for four solar neutrino experiments. From Reference 21.

The SAGE and GALLEX experiments use gallium as the detector and can detect neutrinos whose energies are >0.2 MeV. The average for the gallium experiments is 74 SNU, but a rate of 137^{+6}_{-7} SNU is expected (bar graph on the right in Fig. 6). This conflict is the “third” neutrino problem. (“As far as the theoretical predictions are concerned, several teams have developed computer codes for predicting the current neutrino flux and, when using the same input data, obtain results that agree to within a few percent”) (see Reference 24). Bahcall (21) also concludes “that at least three of the four operating solar neutrino experiments give misleading results or else physics beyond the standard electroweak model is required to change the neutrino energy spectrum (or flavor content) after neutrino production.”

Earlier, Bahcall and Bethe (25) argued that the SSM is valid and that any nonstandard solar model consistent with the K II data predicts a Homestake detection of 4 SNU, which is too high. So if Homestake is correct, then neutrinos are missing. Now, the neutrinos from the p-p chain are electron neutrinos, whereas at least two other flavors of neutrinos, muon and tau neutrinos, are known. If the electron neutrinos are mixed with some other type of neutrino that has a small mass, then they could change their form in transit from the solar core to the detector. This phenomenon, known as “neutrino oscillations,” could be the answer. This and several other possibilities are discussed in Bahcall (21). Finally, non-standard models of the solar interior (mentioned before) can possibly be tested with helioseismological results from SOHO and GONG. They can detect “p-oscillation” modes of the solar interior and possibly “g” modes as well. The latter have amplitudes in the solar interior that are larger than the “p” modes and relate more closely to the neutrino calculations. **[See note added in proof.]**

Radiative and Convective Zones. The energy produced in the solar core by nuclear reactions is initially in the kinetic energy of the charged reaction products and some in the form of gamma rays of \sim MeV energies. Tracking the energy release in the core to its eventual release into space, to give the solar luminosity, is now done by sophisticated computer programs that use the observed properties of the Sun as constraints. According to the SSM mentioned before, the energy released in the core is transported by radiation through the radiative zone, hence its name, by repeated scattering of high-energy photons as they degrade in energy until they reach the bottom of the convective zone $\sim 10^6$ years after the original gamma ray left the core! The radiative opacity of the solar material is an important parameter input to these calculations (see Reference 26 for a basic definition). The energy at this point is transported to the solar surface by the actual movement of hot gas that rises to the surface, a physical process completely different from energy transport by radiation. This process is convection, the rise of hot gas to the photosphere through a hierarchy of cells of different size so that the cooler gas above sinks and there is a net flow of thermal energy to the solar surface. The variations in temperature, density, and pressure as energy moves outward from the core through these zones are shown in Fig. 3, according to recent SSM calculations (14b). (See also Reference 27 for a clear treatment of heat transfer in the convective zone.)

Photosphere. At optical wavelengths, the photosphere defines the Sun as we see it, and much of the activity of the Sun can be seen by features that change in the photosphere, such as sunspots. As just mentioned, all of the energetic

photons from the core have been absorbed by the time they reach the outer boundary of the radiative zone (see Fig. 2), and the energy they carry is finally transported convectively upward to the photosphere where the gas density is too low to sustain convection. The photospheric surface then radiates essentially as a blackbody because it has been heated to an average temperature of ~ 5800 K. A white light image of the full solar disk actually shows a limb darkening. This is understood as a result of the rapid fall in temperature with height, so that at the limb, the cooler gas at a higher altitude radiates at lower intensity than at disc center. There, the visible radiation comes from a greater photospheric depth where the gas is hotter and its radiant intensity is higher. According to this picture, the lower level of the photosphere has a temperature of ~ 6200 K, the upper level ~ 5400 K, and the average disk temperature is ~ 5800 K.

A study of the photospheric absorption spectra, combined with theoretical models, has led to a determination of the composition of the Sun. Our current knowledge on this important topic also involves a study of spectra of solar prominences and sunspots and direct measurements in the solar wind, topics to be covered later. Table 4 lists the abundances of elements up to calcium ($Z = 20$ (28)), although most elements up to thorium ($Z = 90$) have a measurable photospheric value. The value A in the table is the power of 10 of the element's abundance based on reference to a hydrogen abundance of 10^{12} . The solar composition given

Table 4. Recommended Values for Solar Abundances^a

Z	E1	A_{E1}	Comments
1	H	12.0	Reference element
2	He	10.9	Prominence, flare part., solar wind
3	Li	1.0	Photosphere, spot
4	Be	1.1	Photosphere
5	B	2.3	Photosphere
6	C	8.7	Photosphere
7	N	7.9	Photosphere, corona
8	O	8.8	Photosphere, corona
9	F	4.6	Spot
10	Ne	7.7	Corona
11	Na	6.3	Photosphere, corona
12	Mg	7.6	Photosphere, corona
13	Al	6.4	Photosphere, corona
14	Si	7.6	Photosphere, corona
15	P	5.5	Photosphere, corona
16	S	7.2	Photosphere, corona
17	Cl	5.5	Photosphere, spot
18	Ar	6.0	Corona
19	K	5.2	Photosphere
20	Ca	6.3	Photosphere, corona

^aFrom Reference 28.

by “mass fractions” used in the SSM calculations is typically $X=0.71$, $Y=0.27$, and $Z=0.02$ for the hydrogen, helium, and heavier element abundances, respectively (14). An up-to-date treatment of the topics covered here can be found in *Solar Astrophysics* (27).

Granulation. The upper portion of convection cells can actually be observed, as was done at Mount Wilson which showed Doppler shifts of Fraunhofer lines from individual granules. Excellent observations were made by instruments on the Skylab in 1973. These observations show that the visible solar surface has a granular appearance that continually changes; bright granules move upward, as shown by a small blue Doppler shift of certain spectral lines corresponding to a velocity of ~ 1 km/s. The darker portions of this granulation, when viewing the same spectral lines, show a redshift, indicating cooler material that is falling back into the convection zone. The granules, of typical size ~ 1000 km, are in continual motion and have a lifetime of several minutes. Thus, it is believed that this granulation corresponds to the upper layer of the convection zone. But on a larger scale size of $\sim 30,000$ km, a flow pattern is also seen that is similar to the granulation, where similar upward and downward motion of gases occurs at the centers and edges of the cells, respectively. It is believed that these larger features, called supergranulation, are the imprint on the photosphere of larger convective cells deeper in the convective zone (27,29).

Active Regions. Historically, “activity” on the sun was recognized by visually observing sunspots, which changed in number apparently in an apparent cyclic manner, and the bright patches, called faculae, seen close to sunspots when observed near the solar limb. Today, areas on the photosphere, where spots, faculae, other features, called plages, and dark filaments appear together, are called active regions. Transient phenomena such as solar flares and prominences are manifested in these regions, hence the appellation—active. Figure 7 shows an excellent high-resolution image of faculae and a sunspot near the solar limb at a wavelength of 575.5 nm taken at Pic du Midi, France.

The activity of the Sun, which will be discussed further later, is closely related to, and probably due to, solar magnetism that is evident most strongly in sunspots. The magnetic field on the Sun was detected by Hale in 1908, using the Zeeman effect, which was discovered in the laboratory in 1896 by Pieter Zeeman. Hale discovered the sunspot magnetic field in 1912, again using the Zeeman effect. The sunspot field can be as strong as several thousand gauss. The central darkest area of a sunspot is called the umbra, and the surrounding annular ring the penumbra. (See Reference 30 for excellent photographs of all features of active regions.)

Sunspot Number and Magnetic Polarity Cycles. Many terrestrial inhabitants are aware of the relationship between the spots on the Sun and impressive auroral displays and inconveniences such as high-frequency communication disruptions and power outages. Fewer are aware of the roughly periodic, cyclic nature of the number of spots. The sunspot cycle was discovered in 1843 by a German amateur astronomer, Heinrich Schwab, who, after only two decades of observation, stated, “The total number of sunspots has a period of about 10 years.” After ~ 1850 , the Sun was observed out on a regular basis, so the study of the cycle reported by Schwab could be studied in detail. Actually, sunspots often appear in groups that can sometimes last for a 27-day solar rotational



Figure 7. High-resolution image of faculae and sunspots near the limb (upper center), obtained at the Pic du Midi Observatory, France, at a wavelength of 5750 with filter of 100 Å passband. By permission of R. Muller.

period, the time for a solar rotation relative to moving Earth. A convenient way to understand the characteristics of the sunspot cycle is illustrated in Fig. 8. In the lower part of the figure, the sunspot areas are shown as a function of time. This is simply a time plot of the total area of all of the sunspots on the disc at a given time. The cyclic time behavior of the area is very similar to a plot of sunspot number versus time, in which the time between the peaks is ~ 11 years, although the interval could be a few years shorter or longer. A plot of the relative sunspot number, which is actually a running mean, typically peaks a year or two before the area plot. As pointed out by Zirin (30), sunspot area is a more significant indicator of the degree of solar activity than number. At the top of Fig. 8, the latitude on the sun where a spot appears is plotted versus time. This type of plot, known as Maunder's butterfly diagram, shows that as a cycle proceeds, spots appear at lower and lower latitudes; the first spots of a cycle appear near 30° N and S latitudes, and the last ones of the cycle appear near the solar equator. This latitude behavior, named after its discoverer, is called Sporer's law. In this plot,

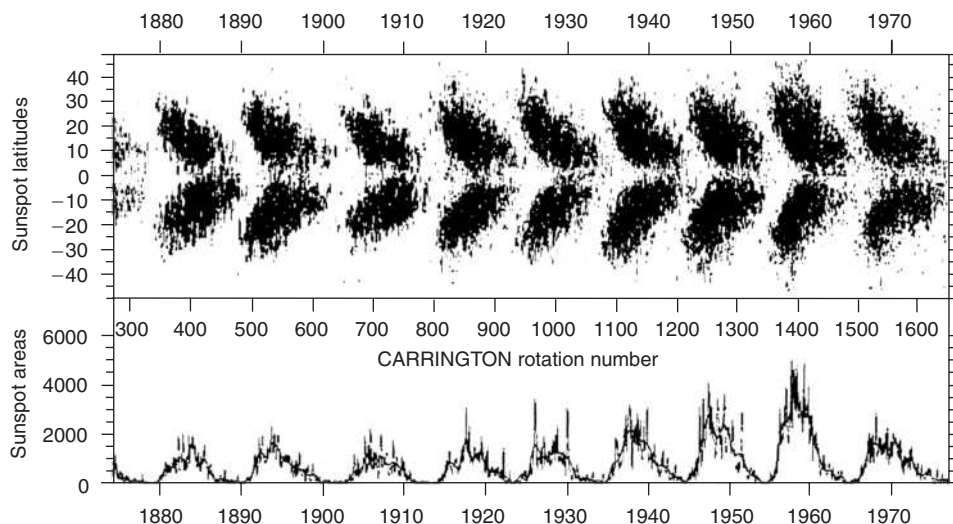


Figure 8. Butterfly diagram (compiled by J.A. Eddy) showing the drift of sunspot position toward the equator. The length of each cycle is about 12 years, causing the old and new cycles to overlap (HAO). From Reference 30.

the length of a cycle is about 12 years and has a small overlap of the end of a cycle and the beginning of the next cycle.

What about the behavior of the magnetic properties of the sunspots? If one considers a pair of sunspots in a given hemisphere during a particular sunspot cycle, the leading spot moving west will have one magnetic polarity (N or S), and the following spot has the opposite polarity, but this pattern is reversed in the other hemisphere. In the next solar cycle, the polarities are also reversed, which means that two sunspot cycles, or 22 years, must elapse before the orientation of the magnetic polarity is the same again. This characteristic of sunspots, discovered in 1912, is known as the Hale–Nicholson law. How are we to understand this fascinating behavior of sunspots?

The most obvious features of the magnetic characteristics of active regions are the strong local magnetic fields, the Hale–Nicholson polarity laws, Sporer’s law, and the 22-year magnetic cycle. Forty years ago, Horace Babcock, who developed the first solar magnetograph at the end of the 1940s with his father Harold, proposed a “solar dynamo” model that explains qualitatively the four features just mentioned (31). The essential ingredients consist of a weak ~ 1 -gauss solar surface field due to an axisymmetric dipole and a rotating Sun that has a convective zone that rotates faster at the equator than at higher latitudes. According to Babcock, it is assumed that the total magnetic flux for the equivalent dipole field, which extends over all space outside the Sun, is $\sim 8 \times 10^{21}$ maxwells (or $\sim 8 \times 10^{15}$ webers) and lies in a thin convective layer about $0.1 R_{\odot}$ thick just below the photosphere. However, the current view is that the dynamo lies in a transitional region between the convective and radiative zone called the tachocline (see, e.g., References 32 and 33).

The differential rotation of the Sun with latitude has recently been determined directly by the Michelson Doppler Imager (MDI) on the SOHO satellite.

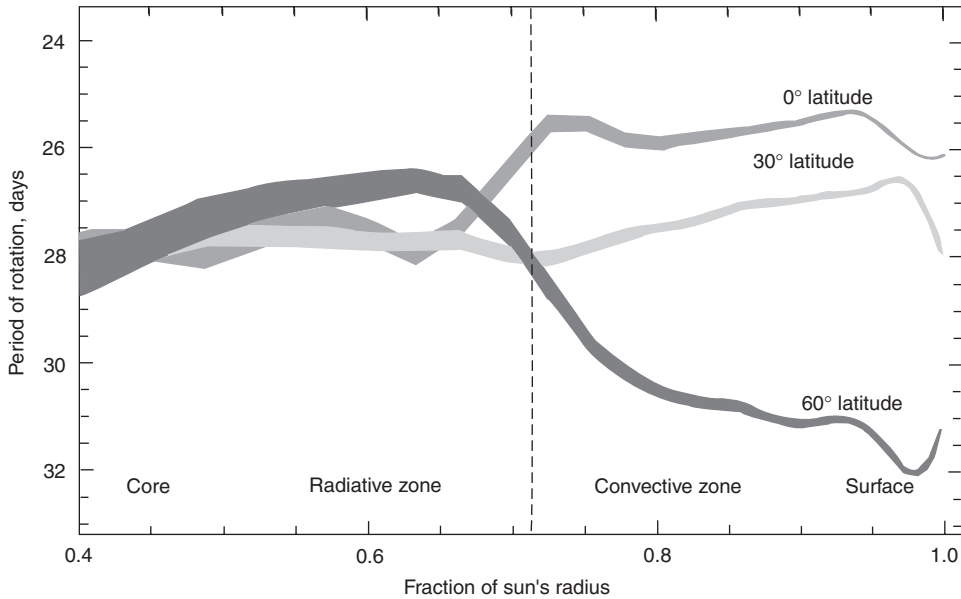


Figure 9. Internal rotational rate of the Sun at latitudes of 0, 20, and 60° has been inferred using data from the Michelson Doppler Imager on the SOHO space craft. Down to the base of the convection zone, the polar regions spin more slowly than the equatorial ones. Below the convection zone, uniform rotation appears to be the norm, although scientists have not yet determined rotational rates within the Sun's core. From Reference 34.

Figure 9 (from Reference 34) shows that the rotational period, as a function of the fractional radius of the Sun, in the radiative zone that extends to $\sim 0.7 R_{\odot}$ from the Sun's center is the same at all latitudes, but in the convective zone, the period varies with latitude. The belief that the convective zone is in differential rotation is strikingly confirmed by MDI measurements. Babcock suggests that if the dipole field is initially normal, about 3 years into a new sunspot cycle, as a result of the frozen-in subsurface field and convective zone differential rotation, "a spiral wrapping of five turns in the North and South Hemispheres (occurs)." The field lines become stretched and twisted leading to a greatly enhanced field up to a factor of 45, depending on latitude. "Twisting... by the faster shallow layers in low latitudes forms 'ropes' with local concentrations that are brought to the surface by magnetic buoyancy to produce bipolar magnetic regions (BMRs) with associated sunspots and related activity." The essence of Babcock's original model is shown in Fig. 10, where the opposite polarities of the BMRs in the Northern and Southern Hemispheres are clearly shown. (Excellent summaries of Babcock's model, which still gives a clear qualitative picture of the magnetic cycle, may be found in References 27 and 29.)

Chromosphere and Transitional Zone. The fortuitous equality of the angular diameters of the Sun and Moon permits astronomers to see the relatively thin (~ 1500 km thick) and cool (~ 4500 K) lower atmosphere of the Sun for a few seconds before and after the total phase of an eclipse. The fact that this layer, the chromosphere, appears reddish in color is the reason that it is sometimes called

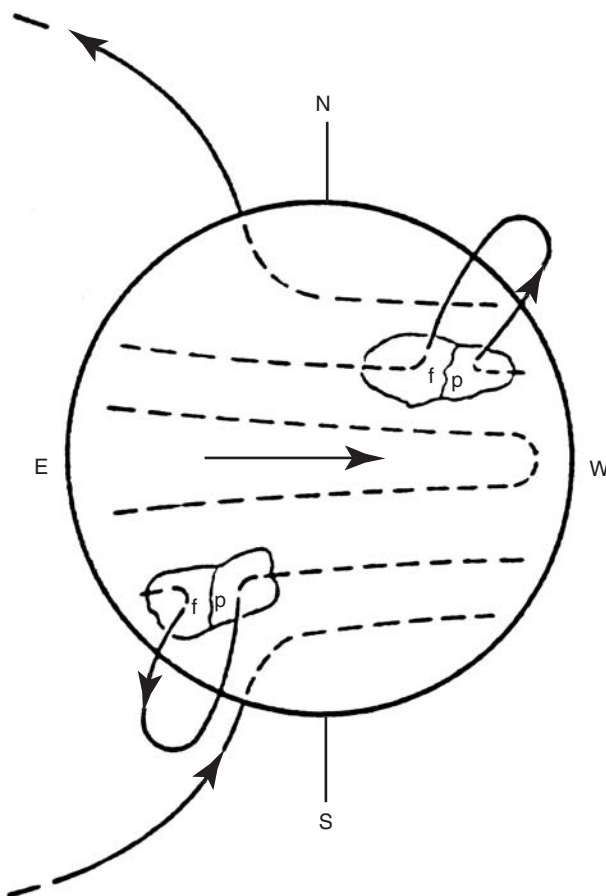


Figure 10. Bipolar magnetic regions (BMRs) are formed where buoyant flux loops of the submerged toroidal field are brought to the surface after several solar rotations have twisted the original global field. The BMRs continue to expand, and flux loops rise higher into the corona. From Reference 31.

the “color-sphere.” The color is due to the Balmer H α red emission line at a wavelength of 6563 Å (656.3 nm). Figure 11, based on model calculations of Vernazza et al. (35), shows a schematic diagram of the temperature versus height above the outer edge of the photosphere.⁴ Thus, the temperature first falls, as expected, for the first 500 km, reaches the “temperature minimum,” and then slowly climbs to ~ 7000 K at an altitude of 2000 km, after which the temperature rises abruptly and reaches $\sim 500,000$ K. The region where the thermal gradient abruptly increases is called the “transition zone.” Beyond the transitional region is the corona where the temperature continues to rise above a million K. Actually the beginning of the chromosphere is defined as just above the “temperature

⁴The “edge” is defined to be at an “optical depth” equal to 1 for a wavelength of 500 nm. Here, a photon will have a $\sim 37\%$ chance of escaping to space.

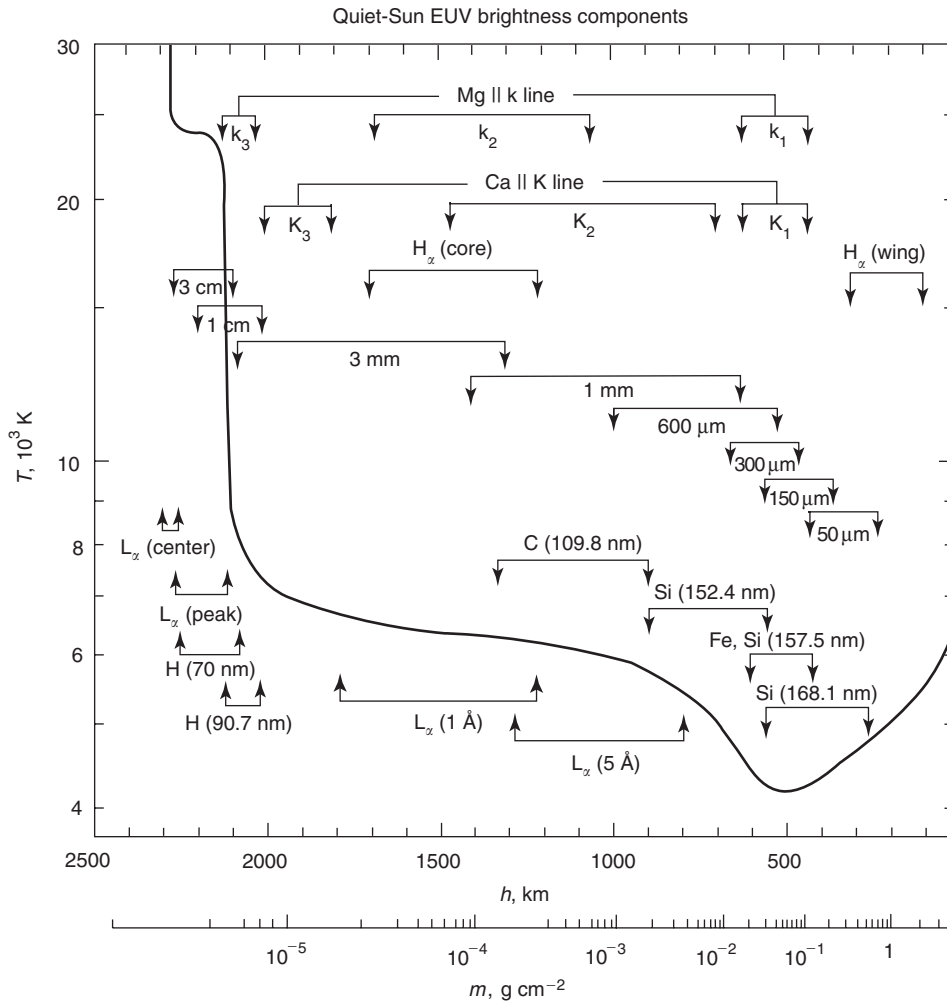


Figure 11. The average quiet-Sun temperature distribution in the chromosphere and transition region derived from the EUV continuum, the L_α line, and other observations. The approximate depths where the various continua and lines originate are indicated. From Reference 35.

minimum” where the temperature rises relatively slowly until the rapid rise. Note that the abscissa is reversed and height increases to the left.

Figure 11 also shows the approximate region in the chromosphere where different emission features arise. Remember that when looking at the solar disc, the strong Fraunhofer absorption features from the photospheric continuum that pass through the chromosphere are dominant in the solar spectrum. Conversely, during a total solar eclipse, viewing the chromosphere at the limb reveals a beautiful spectrum, the “flash spectrum,” devoid of Fraunhofer dark lines. This now becomes an emission spectrum, which enables studying the composition of the chromosphere. Actually the full interpretation of the flash spectrum is very complex and a treatment in depth is given in Zirin (30). At higher altitudes above

the limb, unusual “coronium” lines are seen which were eventually identified as lines from highly ionized Fe and Ca. (These lines are discussed in the next section on the corona.)

When observed in the red $H\alpha$ line of hydrogen, for a few seconds during a solar eclipse, one can see prominences and spike-like jets called spicules. These features are always present above the solar surface but cannot be seen against the bright disc. The spicules can extend up to 10,000 km above the solar limb, have a thickness of about 900 km, and typically move upward at velocities of ~ 25 km/s with lifetimes of ~ 5 min. They are concentrated at the edges of the supergranulation cells. Of course, the immediate question concerning the chromosphere is how it (and the corona) can reach the higher temperatures shown in Fig. 11. The corona can also be observed in the absence of a solar eclipse using coronagraphs on Earth or in space.

Corona. During a total solar eclipse, depending on the phase of the solar cycle, the light that extends well beyond the disc and the narrow chromosphere takes on different forms and is typically as bright as the full Moon ($\sim 10^{-6}$ of the sun’s brightness near the chromosphere falling to $\sim 10^{-8}$ at $\sim 2R_{\odot}$). This “crown” of light is called the solar corona. The most dramatic characteristics of the corona are its high temperature and its extremely low density, surprising for such a luminous object. The changing forms of the corona are associated with the sunspot cycle and the changing magnetic structure. Near the minimum solar sunspot number, the coronal intensity near the poles is depressed relative to that when the solar activity is high.

The corona, observed during an eclipse, emits a continuous spectrum similar to the photospheric continuum but has no Fraunhofer lines and instead bright emission lines, first observed in 1869. It is believed that the continuous spectrum is due to the scattering of the disc continuum from free electrons (Thomson scattering). The lines, it was thought, are from a new element, “coronium.” Then in 1941, Grotrian and Edlen identified the lines as “forbidden” transitions from highly ionized atoms such as Fe X, Fe XIV, and Ca XV, which require million degree temperatures. So the increasing temperature of the solar atmosphere above the temperature minimum reaches a few million K—more than two orders of magnitude hotter than the photosphere. This surprising fact has yet to be explained. Several mechanisms have been suggested, but from 1948 until the late 1970s, sound waves produced by the material motions in the convective zone, it is believed, produce supersonic shock waves that have sufficient energy to heat the corona. However, space-based observations did not detect any shock motions at the lower edge of the corona; presumably, if the shocks are produced outside of the photosphere, they dissipate all of their energy in the chromosphere, explaining the temperature rise seen in Fig. 11. Alternative mechanisms all seem to involve ways of transforming magnetic field energy to heat in a low-density gas, and there is at least circumstantial evidence for an analogous mechanism. This comes from the Einstein Observatory observation of intense X-ray (therefore, hot) coronas around dwarf K and M stars, which may have strong magnetic fields, because large star spots have been detected on their surfaces. Therefore, it would be reasonable to expect that the Sun’s corona could be heated in a similar manner. (A very readable description of attempts to explain coronal heating is found in Reference 29.)

The part of the corona described above that extends to $\sim 2R_{\odot}$ is called the K (for “kontinuerlich”) corona. Beyond $2R_{\odot}$, faint Fraunhofer lines (the F corona) from the photosphere are visible. This is the near-Sun enhanced “zodiacal light” produced by dust in the interplanetary medium.

The corona is now studied in EUV and X-ray wavelengths from space using sounding rockets and satellites. The SKYLAB mission in 1973 obtained soft X-ray images of the inner corona that showed bright loops outlining magnetic field lines that connect opposite magnetic poles of a bipolar magnetic region.

Solar Wind. We understand today that the solar wind (SW) is the dynamic expansion of the Sun’s hot, ionized corona, a plasma that moves radially away from the Sun at a speed of ~ 400 km/s and extends throughout the heliosphere beyond Earth. The study of certain tails of comets led Ludwig Biermann in 1950 (36) to propose “*solare korpuskular-strahlung*” as the explanation for the deflection of their tails⁵ in the radial direction away from the Sun. This proposal credits Biermann with discovering the SW because we now know, from direct measurements in space, that it is composed predominantly of electrons, protons, alpha particles, and other ions from the Sun. The helium abundance of $\sim 4\text{--}5\%$ is apparently enhanced in the corona compared to the photosphere.

The SW has been studied intensively since the first space probes and planetary missions were launched. Its properties are continuously monitored by instruments on several spacecraft beyond Earth’s influence. The SW plasma properties vary with solar activity; the low speed and high speed winds have speeds of 327 and 702 km/s, respectively. For the average SW (speed $\sim 468 \pm 116$ km/s), the average flux of ions, directed radially away from the Sun, is $\sim 4 \times 10^8$ particles/cm²s at 1 AU. The temperatures of the principal SW constituents, protons, electrons, and alpha particles, determined in the frame of the bulk plasma motion, are 120,000, 140,000, and 580,000 K, respectively. The alpha-particle abundance is highly variable in association with transient energetic particle events.

An important feature of the SW, directly associated with the Sun’s large-scale magnetic field, is the variability of its magnetic field. In particular, the radial component of the interplanetary magnetic field, swept out by the SW, switches polarity several times as the solar rotation sweeps the field past Earth in ~ 27 days (the synodic period of the Sun’s rotation, i.e., relative to Earth.) indicating sectors of opposite polarity. The field lines of opposing directions, it is understood, are separated by a thin current sheet, lying approximately in the plane of the ecliptic, which is warped, so as the Sun rotates, the alternating polarities are seen. The sectors of opposite magnetic polarities are referred to as the “sector structure” of the interplanetary magnetic field.

The pioneering work in developing a theoretical model for the SW was done by Eugene Parker (37). See also Parker (38). He assumed that the SW was driven by the thermal pressure of the hot coronal gas. The plasma in the corona is sufficiently hot that it is not gravitationally bound. This mechanism appears sufficient to account for the low-speed SW, but not for the high-speed SW. On the other hand, the origin of the high-speed SW seems to be directly associated with coronal holes. However, the origin of the SW is still not fully understood, and this

⁵The cometary tails referred to here are “ion” tails as opposed to “dust” tails.

remains one of the major areas of investigation in solar physics along with heating of the corona. (A detailed treatment of the solar wind can be found in Reference 27.)

Coronal Holes. Apparently, coronal holes were first observed by X-ray telescopes carried on early sounding rocket flights. In 1973, a telescope on SKY-LAB photographed the full solar disc in soft X rays (< 1 keV). The image obtained (see Reference 39 for an impressive example) showed bright emission over and above most of the disc that extended into the corona. However, near the central meridian, a dark region extended from pole to pole. The impressive feature observed was a large coronal hole, a part of the corona, which is slightly cooler ($\sim 10^6$ K) than the bright hotter $\sim 2\text{--}3 \times 10^6$ K regions. The brighter regions that generally cover the largest area of the corona are due to hot, X-ray emitting gas confined to closed magnetic loops whose foot points are anchored in the photosphere at bipolar magnetic sites or at more complex magnetic sites. The magnetic field lines of coronal-hole regions are open, that is, only one end of a line is tied to the photosphere at a unipolar magnetic site, and the field at several solar radii presumably becomes effectively radial, tied to the solar-wind flow. Using X-ray photographs of coronal holes and measurements of the magnetic fields at the photosphere, the high coronal magnetic field geometry is computer modeled, assuming that there are no electric currents above the photosphere, that is, that the coronal field is potential. When high-speed SW flows are extrapolated back to the Sun along field lines, it is found that the magnetic polarity of the stream in space corresponds to that of the coronal hole at the Sun. So the source of the high-speed SW is a coronal hole!

As mentioned before, the Parker model of the SW assumes that it is driven by thermal pressure, but that is not sufficient to accelerate the fast SW. One possibility is the presence of Alfvén waves that propagate along magnetic field lines and can, if sufficiently intense, carry the plasma along to form the fast SW.

Another point of historical interest is that some geomagnetic disturbances (e.g., cosmic-ray variations) due to enhanced particle fluxes at Earth recurred in a 27-day period. “M regions” on the sun were hypothesized as the source of the corpuscular radiation that causes the disturbance (see Reference 40). These regions are now identified as low-density coronal-hole regions near the sun. (A detailed discussion of this topic and its connection to the SW can be found in Reference 41).

During the first orbit of the Ulysses mission, 1992–1998, Keppler (13) in a recent review reports that there was minimal solar activity so that large polar coronal holes were present and the Solar Wind Isotopic Composition Spectrometer (SWICS) observed fast SW at speeds of 750–800 km/s at the higher latitudes. When Ulysses was at lower latitudes, above the so-called “streamer belt,” the slow SW at speeds of 300–450 km/s was observed, consistent with the view that it arises from closed magnetic regions. Apparently, polar coronal holes extend down to a latitude of 45° near the Sun (within $< 1 R_\odot$ of the photosphere). The Ulysses observations also indicate a continuous mass loss, during a solar minimum, of $\sim 6 \times 10^8$ kg/s that is miniscule compared to the solar mass of $\sim 2 \times 10^{30}$ kg. Above 45° latitude, the fast SW makes the heliosphere extend approximately in the polar directions, but because the Sun’s magnetic dipole axis is $\sim 10^\circ$ from its rotational axis, the extension must wobble in a ~ 24 -day sidereal

period. It is interesting that the boundary between the slow and fast SW is sharp and occurs within $< \sim 2^\circ$ of latitude. The SWICS instrument has also shown that the chemical compositions of the regions are different (42). The Ulysses instruments provide considerable data on the properties of the Interplanetary Medium (IPM), particle acceleration in so-called “co-rotating interaction regions,” and the “anomalous cosmic rays.” Some of these topics are reviewed in Keppler (13). The Ulysses mission will continue until at least 2001, so we can expect many more valuable results related to solar activity and phenomena in the IPM. (See the section on Space Weather, in this article.)

Solar Activity. Solar activity has traditionally referred to dynamic and transient changes in photospheric active regions, where faculae, filaments, plages, and sunspots are seen. Even solar flares, historically the most dramatic symbol of solar activity, result from subtle changes in loop magnetic fields coupled to the photosphere. Within the past 30 years, vast eruptions have been observed in which 20 billion tons of matter have been expelled from the sun in a single event, a coronal mass ejection (CME), that sometimes has disastrous effects on Earth. All of these phenomena result from what might more appropriately be called solar magnetic activity. Great advances have been made in understanding the causes of solar activity, but major new missions will be necessary to unlock the secrets of this aspect of the Sun. We will first discuss some important properties of flares in the next section and then, more briefly, CMEs near the Sun.

Solar Flares and Prominences. In 1859, an intense brightening was observed near a large sunspot group by English observers Carrington and Hodgson; it was the first recorded report of a solar flare, or what is sometimes also called a solar eruption (see Reference 43 for an early history of astronomy). We know now that this type of transient event, which usually can last for minutes or hours, is caused most certainly by the release of energy that resides in the strong magnetic fields that are present in a plage/sunspot region. Before discussing the possible cause of solar flares, some of their general properties must be presented. First, a solar flare is the most energetic *transient* phenomenon that takes place in the solar atmosphere; the largest releases as much as 10^{32} ergs overall, although CMEs may release a comparable amount of energy. This is more than 2 billion times more energy than is released in a 1-megaton nuclear explosion. In addition, it is now known that a large flare emits a major fraction of its energy as electromagnetic radiation (photons), ions, and electrons. The photon energies range from about 10 microelectronvolts (μeV) for meter-wave radio emissions, to electron volts, for optical emissions, and to a billion eV (GeV) for gamma rays. Major flares can produce both ions and electrons that have relativistic kinetic energies, as well as relativistic neutrons, that can reach Earth. The frequency with which flares occur follows the sunspot cycle and, as mentioned earlier, the more flares occur, the larger the total area of sunspots.

Because the intensity of flare emissions appears roughly proportional to the area of the flare in the optical region of the electromagnetic spectrum, solar astronomers have developed a classification scheme to report on the “importance” of a flare. In particular, flares are routinely observed in the light of the red Balmer line of hydrogen, the $\text{H}\alpha$ line whose wavelength is 656.3 nm. Now, the microwave radio flux and the soft X-ray intensity are also routinely measured as well. Table 5 (from Reference 30) shows the classification scheme used to report

Table 5. Flare Classification by Area^a

Area, square deg	Area in $10^{-6} A_{\odot}$	Class	Typical flux at 5000 MHz, sfu	Typical SXR class
≤ 2.0	≤ 200	S	5	C2
2.1–5.1	200–500	1	30	M3
5.2–12.4	500–1200	2	300	X1
12.5–24.7	1200–2400	3	3000	X5
> 24.7	> 2400	4	30000	X9

^aFrom Reference 30.

the properties of an optical flare developed by the International Astronomical Union (IAU) in 1966. The first and second columns in the table refer to the observed area of the enhanced H α emission region in terms of square degrees on the solar surface⁶ in the heliocentric coordinate system (column 1) or in millionths of the area of the solar hemisphere (column 2), and the third defines the optical class of the flare.

Sometimes a letter, f, n, or b, follows the letter class indicating that the estimated visual brightness is faint, normal or bright, respectively. The fourth column gives the microwave flux at a frequency of 5 billion cycles/s (5 GHz) in solar flux units ($1 \text{ sfu} = 10^{-22} \text{ watts/meter}^2/\text{Hertz}$), and the fifth column gives the soft X-ray flux (SXR) in the (1–8 Å) band, as measured by the NOAA GOES spacecraft. The SXR class labels Bn, Cn, Mn, and Xn mean that the X-ray flux is $>n \times 10^{-7}$, $>n \times 10^{-6}$, $>n \times 10^{-5}$ or $>n \times 10^{-4} \text{ watts/m}^2$, respectively. The number after the letter is a multiplier. As an example, if a major flare is designated as 4bX9, its H α emission covers an area of >2400 millionths A_{\odot} , is bright, and has a soft X-ray intensity of $9 \times 10^{-4} \text{ watts/m}^2$ at Earth.

In the flare literature, there are numerous different types of flares depending on their shape or what technique is used for observing, but in H α , the designations compact and two-ribbon seem sufficient (44). Generally, flares occur in magnetic loops or flux tubes whose foot points are rooted in, or near, sunspots. Compact flares are small and apparently have a single loop, whereas two-ribbon flares are much larger and have two bright ribbons moving away from a single dark ribbon or filament. When viewed on the disc, a prominence appears as a dark filament. The larger flares can last for hours, and the brightest are seen as white light flares. The H α flare is often called a chromospheric flare, or “cool” flare, as contrasted to the “hot” flare that occurs in the corona at a temperature of $\sim 10^7$ K. (See an extensive discussion in Reference 45).

It is instructive to see a flare in a wavelength other than in the standard H α band; and both sounding rocket and satellite images are readily available now. On 11 September 1989, a sounding rocket carrying a high-resolution X-ray telescope, NIXT (Normal Incidence X-ray Telescope, the forerunner of TRACE), observed the solar corona in a narrow wavelength band centered at 6.35 nm, which includes X-ray spectral lines from highly ionized Fe XVI and Mg X atoms.

⁶One square degree on the Sun corresponds to an area of $1.48 \times 10^8 \text{ km}^2$. One solar hemisphere has an area of $3.04 \times 10^{12} \text{ km}^2$.

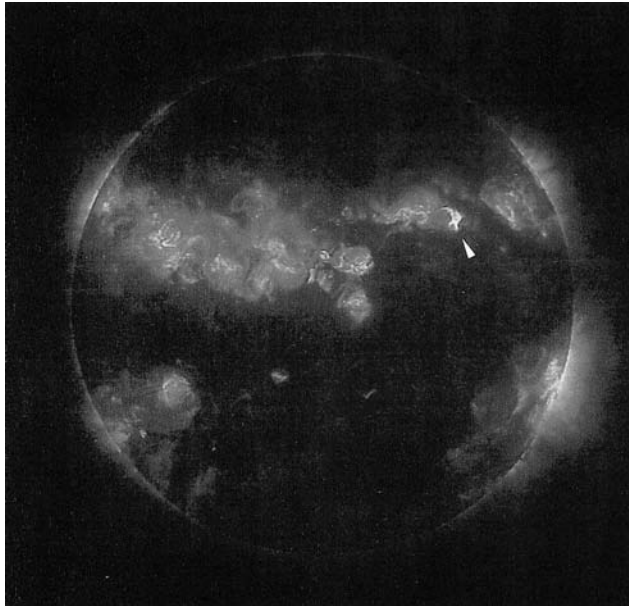


Figure 12. A high-resolution X-ray image of the solar corona that has active regions and a flare (indicated by the arrow). The angular resolution is $0.75''$, and the temperature of the brightest areas is about 3×10^6 K. This image was made by L. Golub et al. in 1989 from a rocket using the 0.25-m X-ray telescope NIXT (Normal Incidence X-ray Telescope) in a narrow wavelength range around 6.35 nm; this range includes the emission lines of Fe XVI at 6.37/6.29 nm and of Mg X at 6.33/6.32 nm. In contrast to an image-forming X-ray telescope of the Wolter type that has grazing incidence, here the image is produced at normal incidence as by an ordinary optical mirror. Reflection of the X rays is made possible by vapor deposition of alternating thin layers of cobalt and carbon, so that constructive interference results for the wavelength 6.35 nm. NIXT was developed by L. Golub and co-workers at the Smithsonian Astrophysical Observatory, Cambridge, together with the IBM Thomas J. Watson Research Center, Yorktown Heights, New York (photo courtesy of SAO and IBM Corp.). (See Reference 46 for description of rocket flight.) This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

The coronal temperatures required to produce these lines is in the range of a few million degrees ($\sim 10^7$ K). The image obtained is shown in Fig. 12 (from Reference 46) and shows the full solar disc where bright emission is across the several active photospheric regions. In addition, the most intense feature, northwest⁷ of the Sun's center is a small two-ribbon subflare, SbC5, so the soft X-ray intensity is 5×10^{-6} W/m² in the 10- to 80-nm band. This flare was observed by NIXT 2 minutes after the GOES monitoring satellite observed the peak of the X-ray emission and, judging from its brightness compared to the coronal emission across the several active regions, its temperature is probably 10^7 K. What is particularly interesting and significant about this observation is that the tangled coronal structure, controlled by the local magnetic fields, is undoubtedly due to similar magnetic complexity in the active photospheric regions.

⁷Generally, in images of the solar disc, east is on the left, and west is on the right.

The basic question of what causes a flare is still unanswered, but optical and magnetic observations of active regions and the morphology of two-ribbon flares have led to the general view that reconnection, or annihilation, of opposing magnetic fields occurs in an “arcade” of magnetic loops that are anchored in the photosphere. It is also now believed that the energy release occurs in the corona and there is adequate energy stored in sunspot magnetic fields so that only a small fraction of the existing fields need be annihilated to supply the energy for a major flare. (See the next section for a specific example—the Masuda flare.)

There is a vast literature on solar flares that was developed before the space age began, and key observations of their electromagnetic emissions were made by both ground-based optical and radio telescopes. Much of the early and current knowledge about flares is expertly treated by Zirin in *Astrophysics Of The Sun* (30) and earlier by Svestka in *Solar Flares* (47). Recent discussions may be found also in Tandberg-Hanssen and Emslie (48) and Somov (49).

High-Energy Flare Emissions. A singular event occurred in 1942, when several cosmic-ray detectors on Earth responded to an increased flux of particles, presumably produced in association with an intense flare that occurred 2 days earlier. If the increased ground-level cosmic-ray flux was due to protons at the top of the atmosphere, then the protons must have relativistic kinetic energies. This “solar flare effect” in ground-level high-energy cosmic-ray detectors recurred about every 4 years until after WWII, when cosmic-ray groups around the world began carrying instruments to high altitudes on balloons and the so-called solar cosmic rays could be studied at lower energies using other instruments. It is this aspect of solar flares that is most important to focus on because high-energy flare emissions produce the greatest effects on Earth and on spacecraft. Recognition of this fact led in 1963 to a major conference on “The Physics of Solar Flares” at the Goddard Space Flight Center (see Reference 50). This article concentrates mainly on recent results obtained from SMM, YOHKOH, CGRO, SOHO, and TRACE after some brief history. (See Reference 11 for a review of results from 1942–1995.)

The dramatic increases in ground-level cosmic-ray intensity are often referred to as ground-level events (GLEs), enhancements of the count rate of a cosmic-ray intensity monitor. The first GLE observations were made using ionization chambers and Geiger–Mueller counters that detect predominantly the charged secondary components produced in the atmosphere by a spectrum of energetic extraterrestrial particles. After the end of World War II, neutron monitors were developed that respond to the secondary nucleonic component produced by the lower energy particles in the spectrum incident at the top of the atmosphere. See Simpson (51) for a detailed discussion of this subject. Using this new technique, on 19 November 1949, a GLE was observed after a large solar flare. The intensity recorded during the maximum of the event was six times the normal value, a relative increase, which was considerably larger than those of earlier bursts.

Major activity in the early part of 1950 was low; then on 26 February 1956, a large GLE occurred. This event became a powerful stimulus to research on high-energy charged particles and photons associated with solar flares. The neutron monitor increase showed that protons that had energies $>15\text{--}30\text{ GeV}$ were produced in close temporal association with the flare optical and radio

emissions. The protons that caused this GLE, it was believed were accelerated by the “Fermi mechanism” in the turbulent magnetic field produced by the chromospheric solar flare.

The first observations of energetic photons from a solar flare were made on 20 March 1958 from a balloon-borne experiment. These detector increases coincided exactly with the maximum intensity of the optical flare (which began 5 minutes earlier) and with an 800 MHz microwave radio burst! Analysis of the individual detector rates indicated that the burst was caused by bremsstrahlung from $\sim 10^{35}$ electrons that had energies of ~ 0.5 MeV.

After the USSR launched Sputnik, (4 October 1957) and the US satellite program was initiated, Morrison (52) published a stimulating paper predicting the fluxes expected for gamma-ray emission from several celestial objects and other phenomena, including solar flares. Also in the post World War II period, the systematic study of radio emission associated with solar flare “eruptions” began. By the mid-1950s, radio spectrographs were observing solar radio bursts. Eventually, interferometers were added, that could give the spatial locations of the sources, and a concrete picture of the cause of the radio phenomena developed. This led in the 1960s to a two-phase model that explained that the radio bursts are due to accelerated ~ 100 -keV electrons followed in large flares by Fermi acceleration of protons and electrons to very high energies; this explained the GLE events and the low-energy, flare-associated protons detected by satellite, balloon, and riometer (relative ionospheric opacity meter) detectors. As will be discussed, this model existed until the 1980s before new observations complicated the picture.

In the decade 1970–1980, the successful series of OGO and OSO satellites continued solar observations. Experiments on SKYLAB as well as rocket and balloon experiments were dedicated to study the Sun. Instruments on High Energy Astronomical Observatories HEAO-1 and HEAO-3, dedicated to cosmic observations, could also record high-energy solar emissions. In the early 1970s, the University California San Diego (UCSD) group’s X-ray spectrometers on the OSO 7 satellite obtained, copious X-ray observations in the range of 2–300 keV.

An auspicious development in this period was the appearance of the complex active region (McMath 11976) on 11 July 1972 that heralded the occurrence of the August 1972 series of major flares that had concurrent intense high-energy emissions and associated terrestrial effects. On 2, 4, and 7 August 1972, major flares from a single active region produced the most intense and long-lasting radiations across the full electromagnetic spectrum observed until then that had associated, intense, long-duration charged particle fluxes in space. Fortuitously, during a 10-minute period in the rising phase of the 4 August flare, the OSO 7 gamma-ray spectrometer recorded, for the first time (53), an intense gamma-ray line and continuum spectrum, that dramatically confirmed the 1958 prediction by Morrison (52) that nuclear reactions in the solar atmosphere during flares could produce a 2.223 MeV line detectable on Earth. The spectrum obtained also gave evidence of nuclear lines at 0.511 MeV, 4.4 MeV, and 6.1 MeV and continuum photons to 10 MeV. The ground-based optical observations of the two-ribbon flares on 4 and 7 August are discussed in detail by Zirin (30).

These observations implied that acceleration of protons above 30 MeV was closely associated with the production of the relativistic electrons that produced the bremsstrahlung continuum. This concept was not taken seriously though, probably because of the 3-minute time resolution of the observations. Therefore, the two-phase model, mentioned before, was still considered valid. Also, based on the published OSO 7 data, Svestka in his 1976 monograph on solar flares (47) proposed for the first time that electrons and ions were accelerated in the “same (one-step) acceleration process, or that the second-step acceleration (giving rise to >30 MeV protons) immediately follows the process of pre-acceleration.”

Due to the launch of the SMM satellite on 14 February 1980 and the Japanese Hinotori satellite on 21 February 1981, the continuing International Sun–Earth Explorer/International Comet Explorer (ISSE 3/ICE) and Interplanetary Monitoring Probe (IMP) operations, and the coordinated ground-based solar observations of the Solar Maximum Year (SMY), an avalanche of new information on solar flares became available. Satellite observations by the hard X-ray and gamma-ray spectrometers on these new missions provided dramatic new diagnostic tools for determining the properties of particle acceleration associated with solar flares.

Within the first few months of flare observations in 1980 by the gamma-ray spectrometer (GRS) on SMM, it became clear that acceleration of ~ 50 -MeV protons and relativistic electrons was simultaneous to within the instrument time resolution of 1–16 s (54), confirming the earlier suggestion of Svestka (47). Note that this observation is contrary to the two-phase acceleration model mentioned earlier.

During an intense limb flare on 21 June 1980, the SMM GRS detected, for the first time, a burst of energetic neutrons at Earth, after a 1-minute-long burst of gamma-ray lines and electron bremsstrahlung that extended to more than 100 MeV in photon energy. This confirmed the 1951 prediction (55) that relativistic protons, accelerated during a solar flare, could produce a flux of high-energy neutrons observable at Earth. Then, during the impressive east-limb flare on 3 June 1982 that produced intense nuclear line emission, it was shown (56) that the photon spectrum above 10 MeV included contributions from meson-decay gamma rays and electron bremsstrahlung that extended to more than 100 MeV. Neutron monitors in Europe also recorded, for the first time, secondary neutrons produced in the atmosphere by primary solar neutrons (57) (see Fig. 13). The protons that resulted from the decay of solar neutrons in space were observed by ISEE 3/ICE and IMP detectors.

Several flares observed by the SMM GRS show rich gamma-ray line spectra, from which it is possible, in principle, to determine the composition of the ambient solar atmosphere where the nuclear reactions occur. Analysis of the spectrum obtained for the 27 April 1981 flare, a long duration gamma-ray flare, indicates that the ambient medium composition differs from that of both the photosphere and the corona and requires an enhanced neon abundance. For this flare, it is suggested that significant gamma-ray line production could take place in the corona. In impulsive flares, most of the line emission is expected to peak in the lower chromosphere, and there is essentially no contribution from the corona. Figure 14 shows the spectrum for this flare calculated by Ramaty (58) that gives the best fit to the observed line spectrum. In general, the SMM GRS

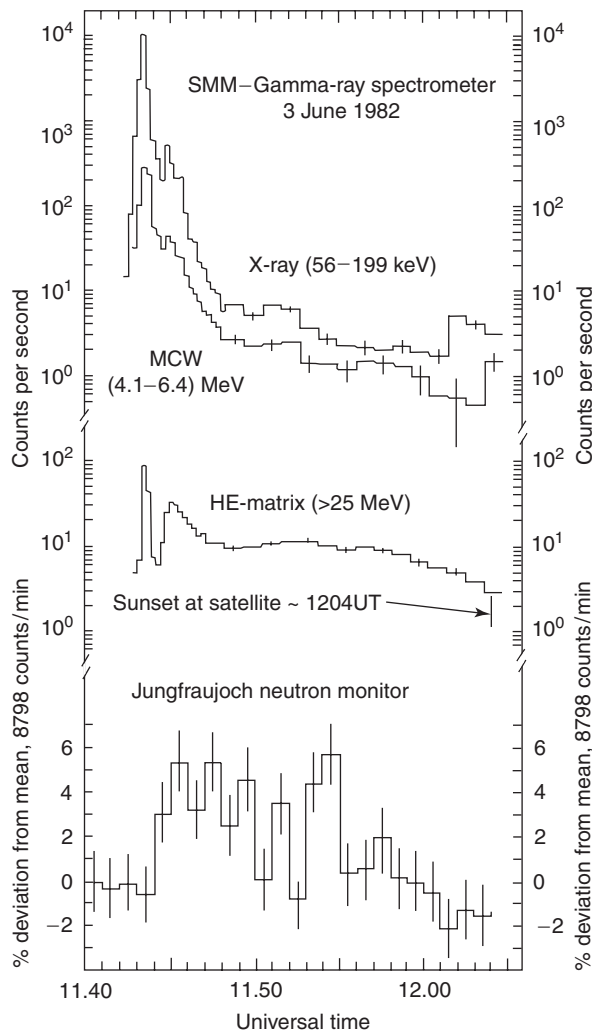


Figure 13. Temporal history for several data channels from the SMM GRS and for the Jungfraujoch neutron monitor count rate for the 3 June 1982 flare. Peak count rates in the GRS (X-ray) and MCW (4.1–6.4) MeV energy bands are uncertain because of pulse pile-up, excessive dead time, and photomultiplier gain shifts and should be used with care. The highest MCW count rates have been estimated using measured live time values, averaged over 16.384 s, and a derived gain shift correction. Error bars are based on count statistics only. From Reference 57.

observations were confirmed by the gamma-ray spectrometer (GRS) on the Hinotori satellite.

By the end of solar activity cycles 20 and 21 in 1986, all of the high-energy radiations predicted (52,59) had been observed; they revealed new insights into the solar flare particle acceleration enigma. Table 6 summarizes the principal characteristics of the emissions from high-energy solar flares. These may be considered observational constraints, which any theory of particle acceleration must satisfy. The temporal evolution of a flare may be

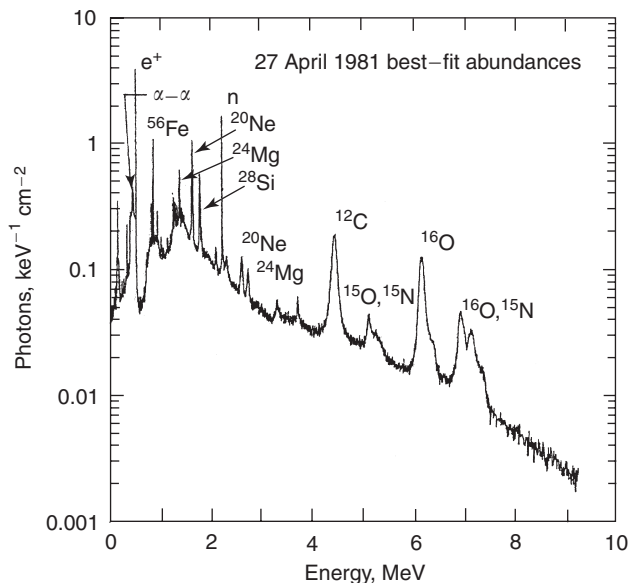


Figure 14. Calculated solar flare gamma-ray spectrum corresponding to abundances that best fit the observed spectrum of the 27 April 1981 limb flare. From Ramaty (58).

separated into three phases: onset, impulsive, and extended phases. A particular flare may be of short duration (a few seconds or minutes) or of long duration (hours); each has emission characteristics shown in the three columns of the table. Table 7 shows the properties of the accelerated electrons and ions of the larger flares.

After the demise of the SMM satellite, on 1 December 1989 the French/Russian satellite, GRANAT, was launched and observed high-energy flare neutral emissions from two flares on 15 May, 24 May, and 11 June 1990. The Russian spacecraft GAMMA-1 provided the only gamma-ray data available on the intense flare at 20:28 UT on 26 March 1991.

Since April 1991, additional solar flare observations were made by the four instruments on the Compton Gamma-Ray Observatory (CGRO). All CGRO instruments directly viewed three X-class flares on 9, 11 and 15 June 1991. Because the instruments on the CGRO, particularly, were much larger than those on the SMM and Hinotori, extended, hours long emissions of gamma rays and neutrons at low flux levels could be identified following some of these flares. The most dramatic example of extended emission was obtained by the CGRO EGRET, following the 11 June 1991 flare when high-energy gamma rays (> 100 MeV) from meson decay and ultrarelativistic bremsstrahlung were observed continuing more than 8 hours after the peak of the flare (60). The PHEBUS spectrometers on GRANAT also recorded gamma-ray line and continuum emissions throughout the intense part of the flare when the EGRET spark chamber and the COMPTEL were disabled because of electronic dead time caused by intense X-ray flux in anticoincidence shields. The EGRET Total Absorption Spectrometer Calorimeter (TASC), however, could also function independently of the spark

Table 6. **Observational Constraints for Acceleration Theory of Temporal Evolution^a**

[I.] Onset phase	[II.] Impulsive burst	[III.] Extended emission
Duration is typically minutes	Sudden enhancement of bremsstrahlung to several hundred MeV	Impulsive events with simple decay
Brightening occurs in UV, soft X rays, or H α	Simultaneous enhancement of γ -ray line emission	Impulsive event and continued production of high-energy emissions
Type III burst at coronal height $\sim (0.4 - 1)R_{\odot}$	Occasional bursts of > 10 MeV photons	Succession of impulsive events and γ -ray line emission dominant over bremsstrahlung
Type I noise storms	Emission of decay γ rays	Succession of impulsive events bremsstrahlung dominant over γ -ray line emission
Increasing intensity of hard X-ray bremsstrahlung	Arrival of high-energy neutrons at Earth $\sim 10 \text{ MeV} < E_n < 2 \text{ GeV}$	
Low-level γ -ray line emission develops	Production of escaping particles (SEP)?	
Bursts of > 10 MeV photons sometimes occur		
New radio sources appear as new energetic emissions appear	Type III/V radio burst	Type II and IV radio bursts

^aFrom Reference 11.

chamber, so high-energy emissions were recorded throughout the flare. After the intense phase of this flare, analysis of the COMPTEL data indicated a flux of neutrons in the energy range of 40–60 MeV, and the EGRET TASC also showed possible evidence of even higher energy neutrons. Figure 15, from Dunphy et al. (61), shows the energy loss spectra in the EGRET TASC for several time intervals during this long event. Phases I-1 and II correspond to the impulsive and extended phases listed in Table 6.

Two important questions posed by the long-term high-energy gamma-ray emission concern the acceleration mechanism/geometry and whether the particles are accelerated in a short impulsive phase that lasts a few minutes, are trapped in magnetic loops, and subsequently precipitate into the lower corona. Or are the particles continually accelerated by processes other than that which produced the particles causing the initial burst of high-energy radiation? The first possibility was considered by Mandzhavidze and Ramaty (62), who

Table 7. **Some Extreme Particle Properties^a**

Parameter	Electrons	Ions (protons)
Number	$10^{41}(> 20 \text{ keV})$ $10^{36}(> 100 \text{ keV})$ $5 \times 10^{34}(> 300 \text{ keV})$	$3 \times 10^{35}(> 30 \text{ MeV})$ $\sim 10^{32}(> 300 \text{ MeV})$
Rise time, s	10^{-2}	> 1
Duration, s	$10 \rightarrow$	$60 \rightarrow$
Total energy, ergs	$10^{34}(> 20 \text{ keV})$ $10^{29}(> 100 \text{ keV})$ $10^{28}(> 300 \text{ keV})$	$10^{30}(> 30 \text{ MeV})$ $3 \times 10^{28}(> 300 \text{ MeV})$
Power, ergs s ⁻¹	$10^{32}(> 20 \text{ keV})$	$2 \times 10^{28}(> 300 \text{ MeV})$

^aFrom Reference 11.

concluded that the trapping model can explain the observations. The second possibility has been considered in interpreting the GAMMA 1 and COMPTEL data for the 15 June 1991 flare. Electric fields produced in a reconnecting current sheet (behind a rising CME) *accelerate the ions and electrons* and explain, particularly, the delayed microwave emission, the 1- to 8-MeV and 100-MeV, and 1-GeV gamma rays. An earlier analysis of essentially the same data concluded that the delayed (gradual phase) emission resulted from prolonged production due to stochastic acceleration. Further interpretations of flares from AR6659 and the question of storage or continuous acceleration of charged particles have recently been reviewed (63).

The high-energy flares discussed before were observed by using nonimaging instruments, so the actual location of gamma-ray and neutron source regions is unknown. However, imaging of relatively small regions on the Sun is possible at photon energies below 100 keV. Due to the launch of the Japanese solar observing satellite YOHKOH and its soft X-ray telescope (SXT) and hard X-ray telescope (HXT), flare observations have been revolutionized. On 13 January 1992, these two telescopes took images of a compact M2.0-class X-ray flare on the west limb. Figure 16 shows a schematic of the flare geometry as deduced from the SXT and HXT images given in Masuda et al. (64). The SXR loop represents the SXT image at $\sim 2 \text{ keV}$, and the three features marked HXR correspond to the major emissions recorded by the HXT above $\sim 25 \text{ keV}$. The flare in X rays below $\sim 33 \text{ keV}$ was of 1-minute duration or less. The image in hard X rays shows the loop top source $\sim 700 \text{ km}$ above the top of the SXR loop, and it is believed that the former is the site of acceleration of the electrons that stream down the two sides of the SXR loop and produce the HXR foot-point sources. The hypothetical magnetic field geometry above the loop top is based on observations of two-ribbon flares

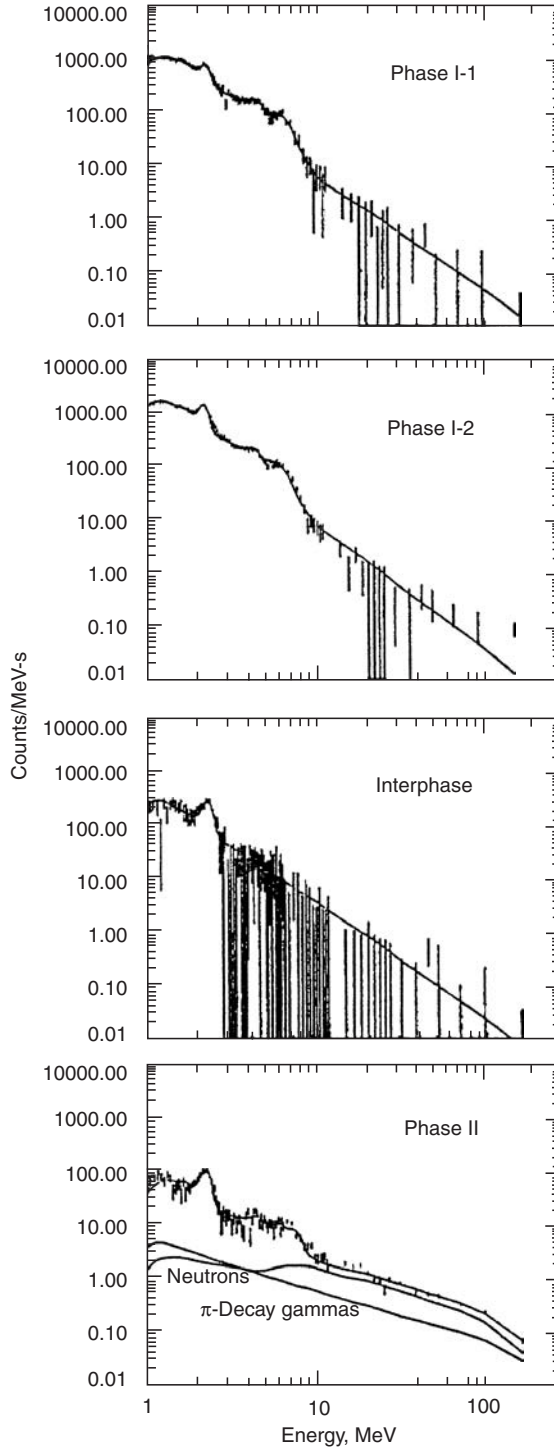


Figure 15. Plots of energy-loss spectra (background subtracted) observed by the TASC for several time intervals during the 11 June flare. The spectra were fit with a multi-component model spectrum described in the text. From Dunphy et al. (61).

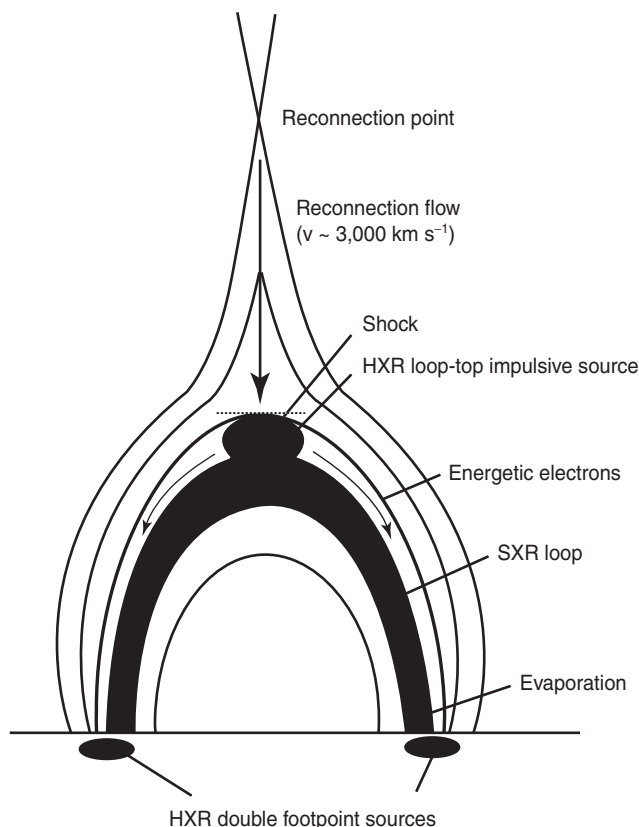


Figure 16. The magnetic-field geometry for reconnection derived from the present observation. The important features are elongated antiparallel magnetic fields above an arcade of closed loops; a current sheet (or neutral sheet) formation between them and the reconnection point impinges on the underlying closed loop and forms a shock, resulting in a high-temperature ($T_e \sim 2 \times 10^8 \text{ K}$) region just above the closed loop. It is also likely that electrons are accelerated in the shock and stream down along the reconnected field toward the double foot-point sources. From Masuda et al. (64).

(65) and is consistent with the view that energy release in a coronal current sheet due to magnetic reconnection produces downward (and outward) moving shocks, which can both heat material in the loop top to temperatures as high as $2 \times 10^8 \text{ K}$ and accelerate nonthermal electrons. Figure 17 shows an SXT image of a major X9 “behind the limb” flare that has major emission above the limb in the corona.

This observation has stimulated further theoretical and interpretive research to understand this type of flare, which may be fundamental for all flares. Recently, Forbes (66) studied possible *trigger mechanisms* for both solar and stellar flares and concluded that it is likely that both “involve the sudden release of magnetic energy” via reconnection but there may be different trigger mechanisms involved simply because the magnetic field structures are probably not the same. In flares, the “trigger” causes a loss of stability, or equilibrium, in a

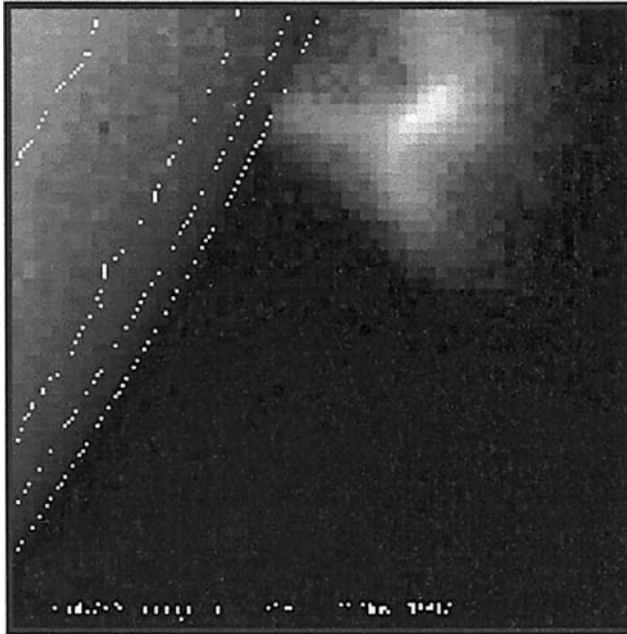


Figure 17. SXT images of the X9 event of 2 November 1992 as the sum of the raw white light and soft X-ray images at about 03:18 UT. The contours show the location of the limb. The flare occurred about 10° behind the limb and was the largest of the Yohkoh era. By Hugh Hudson. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

coronal magnetic field configuration and could be due to new magnetic flux that emerges from the photosphere or twisting of field lines tied at foot points. Either would produce stresses in the field that increase to a limit above which equilibrium is lost.

Coronal Mass Ejections. During satellite missions in the early 1970s, particularly during OSO 7 and the SKYLAB observations, bright transient features that moved rapidly away from the Sun were seen in the corona. These events, which are called coronal mass ejections (CMEs), were observed using white light coronagraphs, which occult the bright disc emission, making an artificial eclipse. They are made visible by scattering, toward the observer, of the photospheric continuum by free electrons (Thomson scattering) from the leading edge of the feature. This technique permits observing the corona from about 2–30 solar radii (R_\odot) from the center of the Sun. CMEs have speeds, measured by the projected position of the leading edge versus time, which range from <50 to ~ 2000 km/s, and estimates of the total mass involved are $\sim 10^{12}$ to 10^{14} kg. CMEs are actual expulsions of matter and magnetic fields from regions lower in the solar atmosphere. Such features are sometimes called coronal transients (see (27)).

The most recent studies of CMEs have found that they are more frequently associated with eruptive prominences than solar flares. The sudden disappearance (or *disparition brusque*) of filaments, it is believed, is associated with most

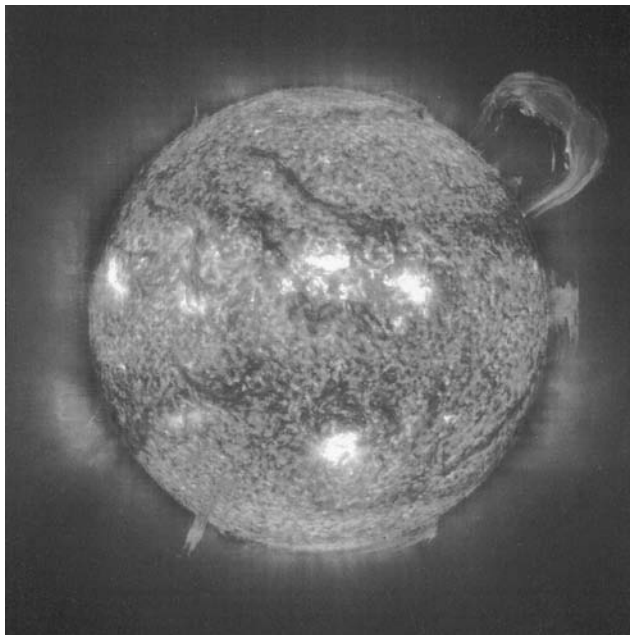


Figure 18. This EIT full Sun image, taken on 14 September 1999 in the He II emission line at 304 Å shows the upper chromosphere/lower transitional region at a temperature of about 60,000 K. The bright features are called active regions. A huge erupting prominence escaping the Sun can be seen in the upper right part of the image. Prominences are “cool” 60,000 K plasma embedded in the much hotter surrounding corona, which is typically at temperatures above 1 million K. If an eruption like this is directed toward Earth, it can cause a significant amount of geomagnetic activity in Earth’s environment and a following spectacular aurora. Instrument: EIT; Taken; 14 Sept 1999, 07:19UT (courtesy of SOHO/EIT.SOHO is a project of international cooperation between ESA and NASA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

CMEs (30). In Figure 18 is shown a recent SOHO image, made by its Extreme-Ultraviolet Imaging Telescope (EIT), of an eruptive prominence on the west solar limb. Presumably the CME that would be associated with such an eruption would lift off from the Sun because the matter is no longer tied to the Sun by the magnetic field of the preexisting prominence. The cause of such eruptions is not known at the present time (30).

Studies have also shown that flares are not always associated with CMEs and sometimes the flare initiation lags the projected time of the CME onset. The CMEs that have speeds in excess of ~ 600 km/s are potential drivers of shock waves moving through the interplanetary magnetic field, and it is believed that acceleration of energetic particles to GeV energies can occur at the shock front. One proposal is that all of the long duration, or energetic gradual, particle events seen at Earth that produce GLEs are caused by particles accelerated by CME shocks. This is actually a very controversial subject because all of the major GLEs at Earth are also associated with major solar flares. As discussed earlier, it is known that GeV ions and electrons are accelerated in a solar flare. Major advances in studying CMEs are currently in progress using the Large Angle

Spectroscopic Coronagraph (LASCO) on the SOHO spacecraft. Further discussion of CMEs may be found in Kahler (67).

Coronae of Solar Type Stars. It might be surprising to find that activity on some solar-type distant stars, ~ 42 light years away compared to the Sun's distance of 8 light-minutes, should be observable from near Earth. The photon flux in any wavelength band would be reduced by a factor of $\sim 10^{14}$ from such a star, compared to the Sun. So-called flare stars have been observed by astronomers (professional and amateur) at visible wavelengths since stellar observations have become commonplace. An important example of flare stars is the binary group known as RS Canum Venaticorum (68), whose discovery is closely related to the discovery of starspots (68). The appellation, "flare stars," was given to this type of star because transient emission was similar to the optical intensity variation (in $H\alpha$) seen in solar flares. Flare stars have also been seen at radio wavelengths since the late 1950s and also at ultraviolet and X-ray wavelengths (see later and also References 69 and 70). In 1974, a rocket observation was made of Capella, an RS CVn variable, by Catura and Acton at soft X-ray wavelengths. This object was identified as a member of "a new class of galactic X-ray sources." It is understood now that the X-ray emission is most likely associated with the degree of magnetic activity on the star, which again is related to the size of starspots on the star.

Since the Einstein Satellite's study of X-ray stars (71) and more recently, using ROSAT data, astronomers have studied the X-ray luminosity of stars ranging in size from giant stars to main sequence and subgiants. Because it is believed that the magnetic activity of the Sun, represented by the number (or total area) of sunspots on the Sun (see Fig. 8), is due to the action of the solar dynamo that results from the differential rotation of the sun, it is believed that the intensity of stellar X-ray emission could be related to the observed rotational speed of the star. This correlation of the X-ray luminosity of a variety of stars with known rotational speeds is shown in Fig. 19 from Pallavicini et al. (71) and has been confirmed in a recent study by Haisch and Schmitt (72). This result also can be understood as follows: recall in our discussion of the origin of sunspots, that two factors were essential: differential rotation of the Sun versus latitude and the presence of a convective zone that provides the conditions for the solar dynamo and leads to the creation of sunspots, emerging magnetic flux, and solar activity in general. Now, the RS CVn binary variables are the most intense emitters of X rays, which it is thought, is related to the short orbital period of the pair about their mutual center of mass. Because the separation of the pair is relatively small, < 0.3 AU, there are strong tidal forces, and the individual star's rotational periods become synchronized with the rapid orbital motion of the pair. It is believed that this is the basic cause of the greater magnetic activity and hence greater X-ray intensity of this system, as shown in Fig. 19. Haisch and Schmitt (72) also discuss the recent 1995 observation of a huge X-ray flare on the binary T Tauri star, V773 Tau. The peak X-ray (0.7–10 keV) luminosity was 10^{33} ergs/s, and the total energy release in the same energy band was 10^{37} ergs; these values are 10^4 and 10^5 higher than the same values for solar flares. T Tauri binary star systems also rotate rapidly, not from tidal action, but because they are young stars that contract from the protostellar cloud stage. It has also been inferred that large sunspots exist in this system, so these observations give

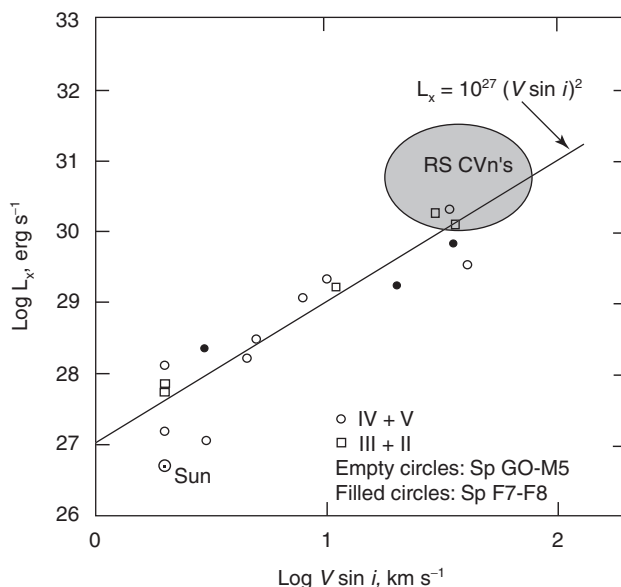


Figure 19. X-ray luminosities plotted versus projected rotational velocities for stars of spectral type F7-F8 (filled symbols) and G0 to M5 (open symbols). The position of RS CVn stars is indicated. From Reference 71.

strong-support to the belief that rotation is the prime cause of their magnetic activity, that is, flaring and this idea carried to the Sun supports the theoretical model of sunspot creation as a result of differential rotation twisting a poloidal magnetic field See also Haisch 1991 (73).

Another important aspect of this solar/stellar connection is the necessity of convective zones for producing a dynamo. The association of coronal activity in the Sun with its magnetic activity, probably produced at the base of the convective zone, is shown in Fig. 20.

Helioseismology. In 1960, Leighton, using the 150-foot tower telescope at Mount Wilson solar observatory to study the Doppler-shift motions of Fraunhofer lines over granulations, discovered that the radial velocities of certain lines oscillated in a period of 5 minutes. These were later explained, independently by R. Ulrich and R. Stein, and another solar physicist, J. Leibacher, by the superposition of nonradial sound waves trapped in the convective zone, which acted as a resonant cavity. Many other modes of oscillation, called acoustic modes, have subsequently been discovered. Because the Sun is a sphere of hot gas, it has oscillatory amplitudes described by spherical harmonics, which are designated by eigenvalues l (angular degree) and m (azimuthal). The 5-minute oscillations now have been identified with l -values from 0 to >1000 . The modes that have small l -values (long wavelengths) can penetrate to the center of the Sun. The waves are also designated as p-modes, or pressure modes, and g-modes, or gravity modes; the latter are due to buoyancy. The g-modes have larger amplitudes and are better for investigating the Sun's structure near the core; however, there is not yet definite evidence of their existence. Clearly, because disturbances are

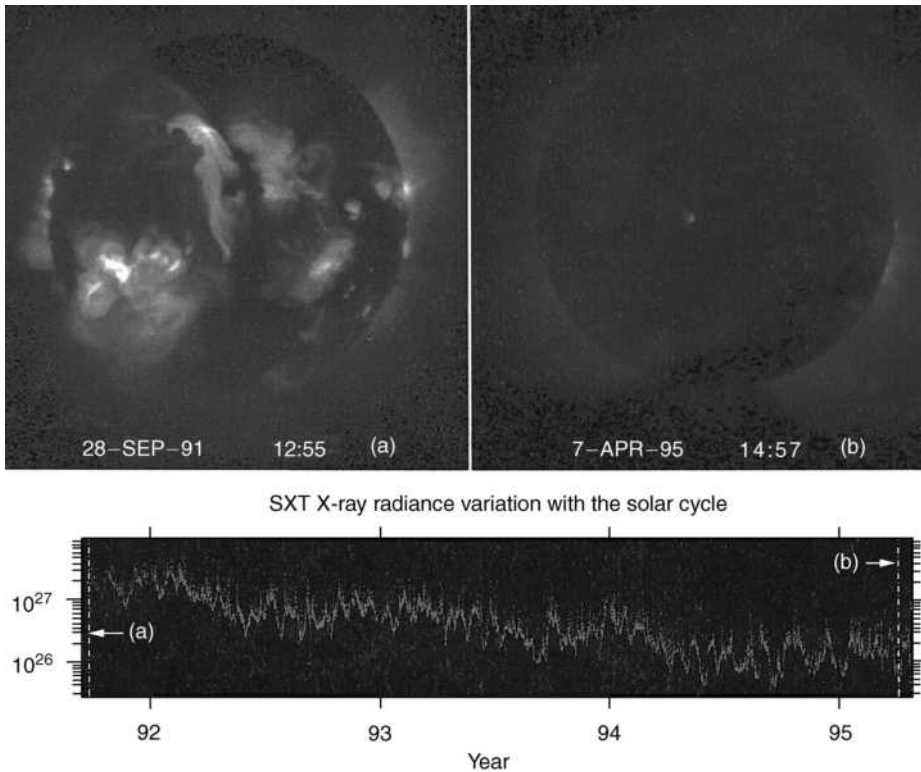


Figure 20. The violently hot solar corona is visible in X radiation. These images from the Yohkoh satellite show how the corona varies with the 11-year solar activity cycle. The bright corona at the left shows high activity, which may be associated with magnetic storms on Earth and other injurious effects. The dark corona on the right shows the present situation (ca. 1995) of low magnetic activity. The plot at the bottom shows the time variation. Activity started to increase again in 1997–1998. The solar X-ray images are from the Yohkoh mission of ISAS, Japan. The X-ray telescope was prepared by the Lockheed Palo Alto Research Laboratory, the National Astronomical Observatory of Japan, and the University of Tokyo with the support of NASA and ISAS. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

propagated at the velocity of sound, which depends on temperature and density, the study of solar oscillations can give information about these parameters in the solar interior.

Currently, the two major projects to study solar oscillations are GONG and the several instruments on the SOHO spacecraft, which were described earlier. The Global Oscillation Network Group (GONG) began operation on 5 October 1995 and consists of six solar observatories, fitted with special telescopes, which are spaced around Earth so at least one will have the Sun in its field of view. SOHO was launched on 2 December 1995, reached the inner Lagrangian point L1 on 14 February 1996, and began constant observations of the Sun. On 25 June 1998, contact with SOHO was lost, but was established again several months later.

The results of new observations from these projects have not been fully synthesized, but one important result gives the sound speed and density profiles throughout the Sun. At a COSPAR meeting in Nagoya in 1998, Shibahashi (74) listed several vital issues in solar physics that helioseismology research will impact, and we mention a few:

- Attempts to solve the solar neutrino problem will benefit by improving the SSM. [See note added in proof.]
- Measuring the rotational rate of the core can aid in determining the evolution of the Sun because of conservation of angular momentum, which suggests an increasing angular velocity with age. Conversely, the surface layers are expected to lose angular momentum to the SW, so the Sun's internal angular momentum distribution should be determined. Related to this is the differential rotational rate of the solar surface and convective zone as a function of latitude.
- The mechanism for depletion of surface lithium abundance is unknown as is the physical cause of the existence of the chromosphere and corona.
- The cause of solar activity and the energy release mechanisms and triggers for flares and CMEs is unknown. How can the change in solar luminosity with solar activity, which is connected to energy transport in the convective zone, be understood?
- Finally, helioseismologists expect to develop a model of the solar interior, independent of the evolutionary SSM.

For further reading on this fascinating subject, see References 27,29,30,74. Of most value are the web home pages:

- GONG: <http://helios.tuc.noao.edu/helioseismology.html>
- SOHO: <http://sohowww.nascom.nasa.gov/>

Space Weather. The effects of solar activity are dramatic in space and even damage or disable satellites. Sometimes the effects on Earth can affect everyday life by disrupting communications or causing electrical power outages. Major effects are caused by intense fluxes of ionizing radiation (primarily energetic protons) and intense X-ray bursts. The most damaging are the large solar energetic proton (SEP) events.

The deleterious effects of solar activity have therefore caused the development of a new *business*—space weather forecasting. The organization charged with this task is the National Oceanographic and Atmospheric Administration (NOAA) in Boulder, Colorado, whose Space Environment Center (SEC) issues daily reports and sends alerts of impending space storms to all who may be affected. Table 8 lists the known “Solar Proton Events Affecting the Earth Environment” which is available on the SEC web page. Following is the SEC web address, where the reader can browse to find topics of particular interest: <http://www.sel.noaa.gov/>.

Notes Added in Proof

1. Firm Evidence for Neutrino Oscillations. This possible solution to the solar neutrino problem was discussed in the original manuscript. Now a new neutrino detector, which began observing solar neutrinos in late 1999, provides compelling evidence that the neutrinos oscillate or change their “flavor” in transit from the solar core to the earth. (See Ahmed et al., *Phys. Rev. Lett.* **89**, 011301 (2002) and *Physics Today*, July 2000).

The detector, The Sudbury Neutrino Observatory (SNO), consists of a kiloton of heavy water in a transparent spherical shell of 12 m diameter and viewed by 9456 photomultiplier tubes. This apparatus is surrounded by a large ultra-pure *light* water shield and resides in a nickel mine at a depth of 6010 m of water equivalent near Sudbury in Ontario, Canada. The SNO can detect the continuous spectrum of ^8B neutrinos above 2.2 MeV (see text—Fig. 4) through the reactions

$$1. \nu_e + d \rightarrow p + p + e^-$$

$$2. \nu_x + d \rightarrow p + n + \nu_x$$

and also by elastic scattering via

$$3. \nu_x + e^- \rightarrow \nu_x + e^-$$

Here x stands for all neutrino flavors. $x = e, \mu, \tau$. Reactions (2) and (3) are sensitive to all neutrino flavors while 1 is only sensitive to the electron-type neutrinos which are directly produced in the fusion reactions in the core of the sun (see text—Table 3). The essential results are

$$\text{flux of } ^8\text{B} \text{ electron neutrinos, } (1.76 \pm 0.11) \times 10^6 \text{ cm}^{-2} \text{ s}^{-1}$$

$$\text{total flux of all neutrinos, } (5.09 \pm 0.062) \times 10^6 \text{ cm}^{-2} \text{ s}^{-1}$$

The SSM prediction for the ^8B flux is $(5.05 \pm 0.91) \times 10^6 \text{ cm}^{-2} \text{ s}^{-1}$, in excellent agreement with the measured total flux of all neutrinos. Thus about two-thirds of the Sun’s ^8B neutrinos have changed their flavor, i.e. presumably transmuted into muon (μ) or tau (τ) neutrinos.

2. RHESSI – A New Satellite to Study Solar Flares. On 5 February 2002, the Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI) was launched by Pegasus XL spacecraft into a low-altitude orbit at an inclination of $\sim 38^\circ$.

The RHESSI mission consists of a single spin-stabilized spacecraft. The only instrument on board is an imaging spectrometer with the ability to obtain high-fidelity color movies of solar flares in X-rays and gamma rays. It uses two new complementary technologies: fine grids to modulate the solar radiation and germanium detectors to measure the energy of each photon very precisely.

RHESSI’s imaging capability is achieved with fine tungsten and/or molybdenum grids that modulate the solar X-ray flux as the spacecraft rotates at

Table 8. Solar Proton Events Affecting Earth Environment January 1988–18 February 2000

Particle event			Particle event		
Start, day/UT	Maximum	Proton flux, pfu at > 10 MeV	Start, day/UT	Maximum	Proton flux, pfu at > 10 MeV
1988			May 13/0910		
Jan 02/2325	Jan 03/0835	92	May 13/0300	May 13/0910	350
Mar 25/2225	Mar 25/2330	58	May 31/1225	Jun 01/0445	22
Jun 30/1055	Jun 30/1140	21	Jun 04/0820	Jun 11/1420	3,000
Aug 26/0000	Aug 26/0045	42	Jun 14/2340	Jun 15/1950	1,400
Oct 12/0920	Oct 12/0930	12	Jun 30/0755	Jul 02/1010	110
Nov 08/2225	Nov 09/0635	13	Jul 07/0455	Jul 08/1645	2,300
Nov 14/0130	Nov 14/0235	13	Jul 11/0240	Jul 11/0450	30
Dec 17/0610	Dec 17/0855	18	Jul 11/2255	Jul 12/0205	14
Dec 17/2000	Dec 18/0150	29	Aug 26/1740	Aug 27/1830	240
			Oct 01/1740	Oct 01/1810	12
			Oct 28/1300	Oct 28/1440	40
			Oct 30/0745	Oct 30/0810	94
1989			1992		
Jan 04/2305	Jan 05/0130	28			
Mar 08/1735	Mar 13/0645	3,500			
Mar 17/1855	Mar 18/0920	2,000	Feb 07/0645	Feb 07/1115	78
Mar 23/2040	Mar 24/0110	53	Mar 16/0840	Mar 16/0840	10
Apr 11/1435	Apr 12/0125	450	May 09/1005	May 09/2100	4,600
May 05/0905	May 05/1000	27	Jun 25/2045	Jun 26/0610	390
May 06/0235	May 06/1045	110	Aug 06/1145	Aug 06/1210	14
May 23/1135	May 23/1350	68	Oct 30/1920	Oct 31/0710	2,700
May 24/0730	May 24/0905	15			
Jun 18/1650	Jun 18/1910	18			
Jun 30/0655	Jun 30/0710	17	Mar 04/1505	Mar 04/1735	17
Jul 01/0655	Jul 01/0720	17	Mar 12/2010	Mar 13/0155	44
Jul 25/0900	Jul 25/1225	54			
Aug 12/1600	Aug 13/0710	9,200			
Sep 04/0120	Sep 04/0510	44	Feb 20/0300	1994	
				Feb 21/0900	10,000

Sep 12/1935	Sep 13/0825	57	Oct 20/0030	Oct 20/0340	35	
Sep 29/1205	Sep 30/0210	4,500				
Oct 06/0050	Oct 06/0825	22		1995		
Oct 19/1305	Oct 20/1600	40,000	Oct 20/0825	Oct 20/1210	63	
Nov 09/0240	Nov 09/0610	43		1996		
Nov 15/0735	Nov 15/0910	71				
Nov 27/2000	Nov 28/1105	380		1997		
Nov 30/1345	Dec 01/1340	7,300	Nov 04/0830	Nov 04/1120	72	W/04 0610
			Nov 06/1305	Nov 07/0255	490	W/06 > 1300
	1990					
Mar 19/0705	Mar 19/2315	950		1998		
Mar 29/0915	Mar 29/1005	16	Apr 20/1400	Apr 21/1205	1,700	W/20 1007
Apr 07/2240	Apr 08/1330	18	May 02/1420	May 02/1650	150	Halo/02 1406
Apr 11/2120	Apr 11/2130	13	May 06/0845	May 06/0945	210	W/06 0829
Apr 17/0500	Apr 17/0655	12	Aug 24/2355	Aug 26/1055	670	NA
Apr 28/1005	Apr 28/1735	150	Sep 25/0010	Sep 25/0130	44	NA
May 21/2355	May 22/0750	410	Sep 30/1520	Oct 01/0025	1,200	NA
May 24/2125	May 25/0115	180	Nov 08/0245	Nov 08/0300	11	?
May 28/0715	May 29/0100	45	Nov 14/0810	Nov 14/1240	310	NA
Jun 12/1140	Jun 12/1700	79				
Jul 26/1720	Jul 26/2315	21		1999		
Aug 01/0005	Aug 01/2015	230	Jan 23/1105	Jan 23/1135	14	NA
			Apr 24/1804	Apr 25/0055	32	Halo/24 1331
			May 05/1820	May 05/1955	14	Halo/ 03 0606
	1991		Jun 02/0245	Jun 02/1010	48	Halo/ 01 < 1937
Jan 31/1130	Jan 31/1620	240	Jun 04/0925	Jun 04/1055	64	NW/ 04 0726
Feb 25/1210	Feb 25/1305	13				
Mar 23/0820	Mar 24/0350	43,000				
Mar 29/2120	Mar 30/0330	20				
Apr 03/0815	Apr 04/1000	52	Feb 18/1130	2000		
				Feb 18/1215	13	W/ 18 0954

–15 rpm. Up to 20 detailed images can be obtained per second. This is sufficient to track electrons as they travel from their acceleration sites, believed to be in the solar corona, and slow down on their way to the lower solar atmosphere.

High-resolution spectroscopy is achieved with nine cooled germanium crystals that detect the X-ray and gamma-ray photons transmitted through the grids over the broad energy range of 3 keV to ~ 17 MeV. Their fine energy resolution of about 1 keV is more than sufficient to reveal the detailed features of the X-ray and gamma-ray spectra, clues to the nature of the electron and ion acceleration processes.

The spinning spacecraft pointing at or near the Sun center provides a simple and reliable way to achieve the rotation required for the HESSI imaging technique. The RHESSI mission has already observed several flares, so considerable new knowledge about solar flares should be obtained in its nominal 2–3 year lifetime.

For the current status of the mission and information about the observed flare events, see the RHESSI homepages at <http://hessi.ssl.berkeley.edu> and <http://hesperia.gsfc.nasa.gov/hessi/>

ACKNOWLEDGMENTS

The author would like to acknowledge Terry G. Forbes, who reviewed an early draft of this article and an anonymous reviewer for valuable suggestions. Discussions on relevant high energy solar flare results with Philip Dunphy are greatly appreciated. Jane Fithian expertly prepared computer files for all color and black and white figures with captions. Rosemary Raynes, Katie Makem and Erica Brown formatted the final manuscript, assisted in numerous tasks and verified references, respectively.

BIBLIOGRAPHY

1. Chaisson, E., and S. McMillan. *Astronomy Today*. Prentice Hall International, London, 1999.
2. Abell, G.O., D. Morrison, and S.C. Wolff. The Sun—an ordinary star. In *Exploration of the Universe*, 5th ed., Saunders, Philadelphia, 1987.
3. Kippenhahn, R. *100 Billion Suns: The Birth, Life, and Death of the Stars*, translated by J. Steinberg. Basic Books, New York, 1983.
4. Menzel, D.H. *Our Sun*, rev. ed. Harvard University Press, Cambridge, 1959.
5. Kiepenheuer, K.O. *The Sun*, translated by A.J. Pomerans. University of Michigan Press, Ann Arbor, 1959.
6. Meadows, A.J. *Early Solar Physics*. Pergamon, Oxford, 1970.
7. Hufbauer, K. *Exploring the Sun-Solar Science Since Galileo*. The Johns Hopkins University Press, Baltimore, 1991.
8. Carruthers, G.R. Sounding rocket experiments, astronomical. In *The Astronomy and Astrophysics Encyclopedia*, S.P. Maran and C. Sagan (eds). Van Nostrand Reinhold, New York, 1992, p. 650.

9. Brueckner, G.E. Solar physics, space missions. In *The Astronomy and Astrophysics Encyclopedia*, S.P. Maran and C. Sagan (eds). Van Nostrand Reinhold, New York, 1992, p. 646.
10. Kundu, M., and B. Woodgate. (eds). *Energetic Phenomena on the Sun*. NASA-(STIB)-Conf. Pub. 2439, Washington, DC, 1986.
11. Chupp, E.L. Evolution of our understanding of solar flare particle acceleration: (1942–1995). In *High Energy Solar Physics*, Woodbury, NY, AIP Conf. Proc 374: 3, 1996.
12. Lord, D.R. Space Lab. NASA (STID), Washington, DC 1987, p. 374.
13. Keppler, E. Ulysses finishes its first revolution around the Sun. *Naturwissenschaftler* 85: 467 (1998).
- 14a. Bahcall, J.N., and R.K. Ulrich. *Ap. J.* 170: 593 (1971).
- 14b. Bahcall, J., and H. Pinsonneault. Standard solar models with and without helium diffusion, and the solar neutrino problem. *Rev. Mod. Phys.* 64: 885 (1992).
15. Lang, K.R. *Sun, Earth and Sky*. Springer, Berlin, 1995.
16. Chandrasekhar, S. *An Introduction to the Study of Stellar Structure*. Dover, New York, 1939, p. 453.
17. Bethe, H.A. Energy production in stars. *Phys. Rev.* 55: 434 (1939).
18. Davis, R., Jr. Attempt to detect the antineutrinos from 2 nuclear reactors by the $\text{Cl}^{37}(\bar{\nu}, e^-)\text{A}^{37}$ reaction. *Phys. Rev.* 97: 766 (1955).
19. Bahcall, J., and R. Davis. An account of the development of the solar neutrino problem. In *Essays in Nuclear Physics*. Cambridge University Press, Cambridge, 1982, p. 243.
20. Bahcall, J.N., and G. Shaviv. *Ap. J.* 153: 113 (1968).
21. Bahcall, J.N. Solar neutrinos: Solved and unsolved Problems. in J.N. Bahcall, and J.P. Ostriker (eds). *Unsolved Problems in Astrophysics*. Princeton University Press, Princeton, NJ, 1997, p. 195.
22. Davis, R., Jr., D.S. Harmer, and K.C. Hoffman. Search for neutrinos from the Sun. *Phys. Rev. Lett.* 20: 1205 (1968).
23. Davis, R. A review of the Homestake solar neutrino experiment. In *Progress in Particle and Nuclear Physics*, 32: Elsevier Science, Oxford, 1994, p. 13.
24. Bahcall, J.N. *Neutrino Astrophysics*. Cambridge University Press, Cambridge, England, New York, 1989.
25. Bahcall, J., and H.A. Bethe. A solution of the solar-neutrino problem. *Phys. Rev. Lett.* 65: 2233 (1990).
26. Eddington, A.S. *The Internal Constitution of Stars*. Dover, New York, 1926, p. 108.
27. Foukal, P. *Solar Astrophysics*. Wiley, New York, 1990.
28. Engvold, O. The solar chemical composition. *Physica Scripta* 16: 48 (1977).
29. Noyes, R.W. *The Sun—Our Star*. Harvard University Press, Cambridge, 1982.
30. Zirin, H., *Astrophysics of the Sun*. Cambridge University Press, Cambridge, Cambridgeshire, New York, 1988.
31. Babcock, H. The topology of the Sun's magnetic field and the 22-year cycle. *Ap. J.* 133: 572 (1961).
32. Gough, D. New from the solar interior. *Science* 287: 2434 (2000).
33. Howe, R., J. Christensen-Dalsgaard, F. Hill, et. al. Dynamic variations at the base of the solar convection zone. *Science* 287: 2456 (2000).
34. Lang, K.R. SOHO reveals the secrets of the Sun. *Sci. Am.* 276: 40 (1997).
35. Vernazza, J.E., E.H. Avrett, and R. Loeser. Structure of the solar chromosphere. III. Models of the EUV brightness components of the quiet Sun. *Ap. J. Suppl.* 45: 635 (1981).
36. Biermann, L.F. Kometenschweife und solare korpuskular-strahlung. *Z. Ap.* 29: 274 (1950).
37. Parker, E.N. Dynamics of the interplanetary gas and magnetic fields. *Ap. J.* 128: 664 (1958).

38. Parker, E.N. *Interplanetary Dynamical Processes*. Interscience, New York, 1963.
39. Eddy, J.A. A new Sun – The solar results from SkyLab. NASA(STIB), Washington, DC, 1979, p. 98.
40. Bartels, J. Terrestrial-magnetic activity and its relations to solar phenomena. *Terrestrial Magn. Atmos. Elec.* 37: 1 (1932).
41. Withbroe, G.L. Origins of the solar wind in the corona. In *The Sun and Heliosphere in Three Dimensions*, R.G. Marsden (ed.). Reidel, Dordrecht, 1986, p. 19.
42. Gloeckler, G.J. Geiss, H. Balsgier, et al. The solar wind ion composition spectrometer. *Astron. Astrophys. Suppl. Ser.* 92: 267 (1992).
43. Clerke, A.M. *A Popular History of Astronomy during the 19th Century*. Adam and Charles Black, Edinburgh, 1885, p. 205.
44. Pallavicini, R. The role of magnetic loops in solar flares. *Philos. Trans. R. Soc. A* 336: 389 (1991).
45. Pneuman, G.W. Two ribbon flares: Post (flare) loops. In *Solar Flare Magnetohydrodynamics*, E.R. Priest (ed.). Gordon and Breach, New York, 1981, Vol. I. p. 379.
46. Golub, L., M. Herant, K. Kalata, et al. Subarcsecond observations of the solar X-ray corona. *Nature* 344: 842 (1990).
47. Svestka, Z. *Solar Flares*. Reidel, Dordrecht, Boston, 1976.
48. Tandberg-Hanssen, E., and A.G. Emslie. *The Physics of Solar Flares*. Cambridge University Press, New York, 1988.
49. Somov, B.V. *Physical Processes in Solar Flares*. Kluwer, Dordrecht, 1992.
50. Hess, W.N. (ed.). *AAS-NASA Symp. Phys. Solar Flares*. NASA (STID), Washington, DC, 1964.
51. Simpson, J.A. Evidence for a solar cosmic-ray component. in *The Sun*, G.P. Kuiper (ed.). University of Chicago, Chicago, 1953, p. 715.
52. Morrison, P. On gamma-ray astronomy. *Nuovo Cimento* 7: 858 (1958).
53. Chupp, E.L., D.J. Forrest, P.R. Higbie, et al. Solar gamma ray lines observed during the solar activity of August 2 to August 11, 1992. *Nature* 241: 333 (1973).
54. Forrest, D.J., and E.L. Chupp, et al. Simultaneous acceleration of electrons and ions in solar flares. *Nature* 305: 291 (1993).
55. Biermann, L., O. Haxel, and A. Schluter. Neutrale ultrastrahlung von der Sonne. *Naturforsch.* 6a: 47 (1951).
56. Forrest, D.J., W.T. Vestrand, E.L. Chupp, et al. Neutral pion production in solar flares. *Proc. 19th Int. Cosmic Ray Conf.* 4: 146 (1985).
57. Chupp, E.L., H. Debrunner, E. Fluckiger, et al. Solar neutron emissivity during the large flare on 1982 June 3. *Ap. J.* 318: 913 (1987).
58. Ramaty, R. Flare physics at high energies. In *Astrophysics from the Moon*, M.J. Mumma and H.J. Smith (eds). AIP, New York, 1990, p. 122.
59. Lingenfelter, and Ramaty. High-energy nuclear reactions and solar flares. In *High Energy Nuclear Reactions in Astrophysics*. B.S.P. Shen (ed.). Benjamin, New York, 1967, p. 99.
60. Kanbach, G.O., D.L. Bertsch, C.E. Fichtel, et al. Detection of a long-duration solar gamma-ray flare on June 11, 1991 with EGRET on COMPTON-GRO. *A. A. Suppl.* 97: 349 (1993).
61. Dunphy, P.P., E.L. Chupp, D.L. Bertsch, et al. Gamma rays and neutrons as a probe of flare proton spectra: The solar flare of 11 June 1991. *Sol. Phys.* 187: 45 (1999).
62. Mandzhavidze, N., and R. Ramaty. Gamma rays from pion decay: Evidence for long-term trapping of particles in solar flares. *Ap. J. Lett.* 396: L111 (1992).
63. Hudson, H., and S.M. Ryan. High-energy particles in solar flares. *Ann. Rev. Astron. Astrophys.* 33: 239 (1995).
64. Masuda, S., T. Kusugi, H. Hara, et al. A loop-top hard X-ray source in a compact solar-flare as evidence for magnetic reconnection. *Nature* 371: 495 (1994).

65. Haisch, B.M, K.T. Strong, and M.A. Rodono. Flares on the Sun and other stars. *Ann. Rev. Astron. Astrophys.* 29: 275 (1991a).
66. Forbes, T. Solar and Stellar Flares. *Philos. Trans. R. Soc. A* 258: 711 (2000).
67. Kahler, S.W. Solar Activity, Coronal Mass Ejections. In *The Astronomy and Astrophysics Encyclopedia*. Van Nostrand, New York, 1992, p. 631.
68. Hall, D.S. Binary Stars, RS Canum Venaticorum Type. In *The Astronomy and Astrophysics Encyclopedia*. Van Nostrand, New York, 1992, p. 74.
69. Hjellming, R.M. Stars, Radio Emission. In *The Astronomy and Astrophysics Encyclopedia*. Van Nostrand, New York, 1992, p. 780.
70. Byrne, P.B. Stars, Red Dwarfs and Flare Stars. In *The Astronomy and Astrophysics Encyclopedia*. Van Nostrand, New York, 1992, p. 783.
71. Pallavicini, R., L. Golub, R. Rosner, et al. Relations among stellar X-ray emission observed from Einstein, stellar rotation and bolometric luminosity. *Ap. J.* 248: 279 (1981).
72. Haisch, B., and J. Schmitt. The solar–stellar connection. In *Sky and Telescope*. Sky, Cambridge, 1999, p. 46.
73. Haisch, B.M. Solar–stellar connection. In *The Many Faces of the Sun – A Summary of the Results from NASA's Solar Maximum Mission*, K.T. Strong, J.L.R. Saba, B.M. Haisch, and J.T. Schmelz (eds). Springer, New York, 1991, p. 481.
74. Shibahashi, H. What the helioseismic results mean for the solar interior. In *Helioseismology and Solar Variability – Advances in Space Research*, C. Frohlich and B.H. Foing (eds). *Adv. Space Res.* 24 (2): 137 (1999).

EDWARD L. CHUPP
University of New Hampshire
Durham, New Hampshire

U

U.S. MANNED SPACE FLIGHT: MERCURY TO THE SHUTTLE

Introduction

There are many publications relating to Mercury, Gemini, Apollo, Skylab, Apollo-Soyuz Test Project, and the Space Shuttle. This paper presents an objective review of some of the technical considerations of the design, development, and operation of these spacecraft. Although there were many triumphs, a number of surprises, and not a few setbacks during the 39-year period covered, such incidents will be mentioned only in the context of technical significance. A little more than 19 years elapsed between 20 February 1962, when John Glenn first rode Mercury into orbit, and 12 April 1981, when John Young and Robert Crippen inaugurated orbital flight for the Space Transportation System, or the Space Shuttle as it is commonly called. A comparison of the Mercury capsule and the spaceship Columbia reviews the extent of progress made during these first years of manned flight.

During the early 1920s after approximately 19 years of development, the aviation industry was also entering the transportation phase. But there is a significant difference between the development of airplanes and the development of manned spacecraft. Early airplanes were relatively cheap and easy to build. Often, only one person and at most just a handful would be sufficient to do the entire design job. Consequently, a great number of airplanes was built and flown. This led to a rapid evolutionary process for both the design and the operation of aircraft, and a number of accidents and fatalities provided a strong cooperative influence on any misdirection. The philosophical basis for the design and operation of spacecraft has had to progress without this impartial and unerring guidance of "survival of the fittest." As a substitute, we have had to rely on

intensive analysis, extensive testing, imperfect simulations, and the judgment and experience of the hundreds of key people working on these programs.

American manned spacecraft by type together with a number of weights, dimensions, and features that characterize them are listed in Appendix A. Mercury was a first of a kind design, and the Gemini and Apollo command and service module (CSM) followed in an evolutionary trend. The Apollo lunar module (LM) and the Space Shuttle, on the other hand, represent new and distinctive designs and have unique features for which no prior art existed.

Project Mercury

The National Aeronautics and Space Administration (NASA) was established in October 1958, and within a few weeks, the first manned spacecraft program, Project Mercury, was initiated (1). To execute the program as quickly as possible, it was deemed desirable to choose an existing rocket for the launch vehicle. The Atlas, considered the most powerful that would be available during the desired time period, was chosen. Conservative estimates indicated that the Atlas could orbit approximately 900 kg (2000 lbm) of payload. The spacecraft grew almost 50% during its development period. When finally developed, the entry weight of the Mercury spacecraft was in excess of 1180 kg (2600 lbm). Fortunately, the Atlas performance also grew sufficiently during this period.

The basic purpose of Project Mercury was to expose several test pilots to orbital flight and to have them evaluate the experience so that future and more substantive programs could be planned (Appendix B). To move the program rapidly and because of severe weight constraints, the design of all systems was as simple as possible and provided only the basic necessities of launch, a short orbital flight, and safe descent and landing. A simple ballistic configuration designed to minimize reentry heating was chosen. Although kept as simple as possible, each active system had at least one level of redundancy. Descent was initiated by firing three solid retrorockets; however, a safe reentry was ensured if only two of these rockets fired. A parachute system that was backed up by a reserve system of identical design provided a safe landing. Redundant sets of hydrogen peroxide monopropellant reaction control jets were used. To guard against electrical failure, one set of jets had its electrically actuated valves backed up by a set of mechanical valves directly linked to the astronaut's rotational hand controller. Life support was provided by a pure-oxygen cabin atmosphere at one-third sea-level pressure. This was backed up by a pressure suit that would automatically inflate if the cabin atmosphere dropped to less than one-fourth sea-level pressure. Redundant voice, telemetry, and command signals could be sent and received on a number of communication channels. The Mercury capsule also carried both C-band and S-band radar beacons to assist the radars at the various ground stations that tracked the flight. Finally, power was supplied by redundant battery sets that had independent buses. Heat was dissipated from the Mercury crew compartment by a water evaporator. Figure 1 shows the Mercury spacecraft and its components.

The concept for flight control was very simple. Basically, the Mercury spacecraft was inserted into a low Earth orbit using the launch vehicle guidance

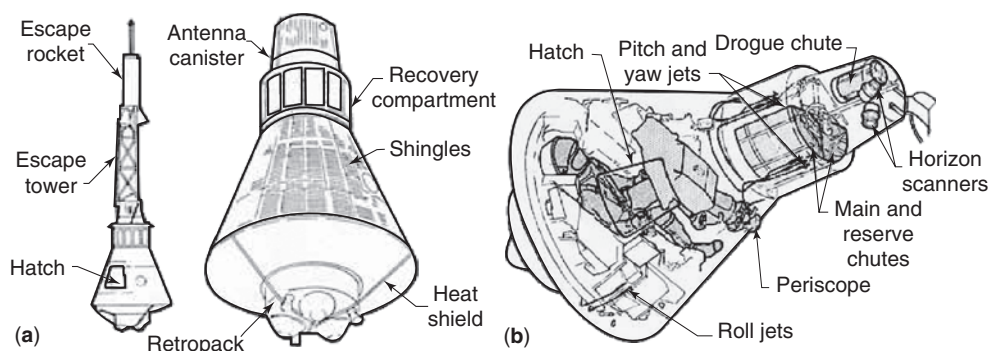


Figure 1. The Mercury Spacecraft and some of its component parts. The escape tower and rocket is shown in picture (a), the heat shield and the retrorocket pack is shown in (b) and the position of the astronaut in the space capsule. These two figures can also be seen at the following websites: <http://www.howstuffworks.com/gif/mercury-capsule-drawing.jpg> and <http://www.howstuffworks.com/gif/mercury-capsule-drawing2.jpg>. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

system. Once in orbit, no further velocity change maneuvers were required until it was time for descent. Return to Earth was accomplished by maneuvering the spacecraft to a retrorocket firing attitude where the heat shield was forward and then firing a cluster of three solid rockets strapped to the heat shield. This deflected the flight path to one that entered Earth's atmosphere. The spacecraft was not designed to produce lift, so it followed a highly predictable ballistic entry trajectory. Consequently, the time when the retrorockets were fired determined the location of splashdown in the ocean. It was recognized that there would be a fairly large dispersion about the planned landing location. However, consideration of emergency descent or aborts during launch that could result in a landing anywhere along the flight track made survival for a period of time on the water after landing and extensive use of location aids a basic design requirement anyway.

Flight control equipment on board the spacecraft was needed only for attitude control, particularly when firing the retrorockets. These functions could be done both manually and automatically. When in automatic flight, an autopilot and two horizon scanners were set up to maintain the vehicle in a fixed attitude with respect to local vertical. An onboard timer also initiated the retrofiring sequence after commanding the optimum attitude for the maneuver. This timer, which was started at liftoff, could be corrected during flight by command signals from the ground. The astronaut could also control attitude with a hand controller either by using an attitude indicator on the instrument panel or by looking through the window. He could also override automatic initiation of the retrorocket sequence. The astronaut could crudely determine his position in orbit by comparing his view of Earth using a clock-driven replica of Earth's globe.

The mission was controlled from the ground. Communications with the spacecraft and tracking data were obtained from a network of stations along the path of the first orbits. There were 16 different stations located to allow a maximum of 10 minutes without communication contact. The data from these stations were sent to the Mercury Control Center at the launch site in Florida. On

the basis of these processed tracking data, the orbital ephemeris was determined. The location of intended splashdown was predetermined to accommodate post-retrofiring tracking and thereby to enhance final location for recovery.

During Project Mercury, six tests were made, two suborbital on a Redstone launch vehicle and four orbital on the Atlas. In preparing for these tests, 17 unmanned flight tests were made; four were made with primates aboard the spacecraft. Only 10 of these 17 tests were successful. In addition to the biomedical studies, these tests provided engineering studies of all of the modes of flight such as escape system operation, entry flight, and performance of the various onboard systems. Of the four manned orbital tests, each was of increasing duration starting with the three orbits made by John Glenn and ending with the 22-orbital flight made by Gordon Cooper. All U.S. manned spaceflights are summarized in Appendix B.

Gemini Program

Based on the success of Mercury and undisputed evidence confirmed by the Russian program that people could function comfortably and effectively in space, it became obvious that the next step would be to develop the role of people in space. Consequently, two new programs were established during 1961 to this end: Gemini, to explore the possibilities and limitations of manned operation in space, and Apollo, a bold program committed to exploration of the Moon and sending people there. In Gemini, the approach was, quickly and, where possible, directly, to capitalize on what was learned in the Project Mercury (2). However, vastly improved capability was desired to explore the feasibility of the following operations: maneuvering in orbit, rendezvous and docking with another vehicle, extravehicular activity (EVA) by the astronauts in pressure suits, guided flight during entry to precisely designated target areas, and establishing and refining flight operation methods. All of these objectives were met with highly positive results that were applied extensively in the Apollo Program and contributed greatly to the success of that program.

The Gemini Program employed the new Titan II launch vehicle. This booster had roughly twice the capability of the Atlas and made it possible for the Gemini spacecraft to carry a crew of two and to incorporate a great number of design features requisite to the program requirements. The Gemini configuration was basically a scaled-up version of the Mercury vehicle. However, the center of mass of the Gemini spacecraft was offset from its centerline. This configuration caused the spacecraft to trim at a slight angle of attack, which produced a small aerodynamic force (lift) normal to the drag force. Although the lift-to-drag ratio (L/D) was only slightly greater than 0.1, it was more than sufficient to steer the entry path to a desired touchdown point. Steering was done by controlling the roll attitude of the spacecraft and thereby the lift vector to deflect the reentry path (up, down, right, left) in the desired manner.

Gemini carried its adapter into orbit. This was the conical structure that connected the Gemini spacecraft to the booster. In this adapter section, which was jettisoned before entry, a great number of systems that were used during orbit were carried. A set of bipropellant rocket motors provided for both

translational and rotational maneuvers. Fuel cells and their hydrogen and oxygen supply tanks were installed in the adapter as were radiators that provided for heat dissipation. The fuel cells and radiator made mission durations as long as 2 weeks possible. The Gemini spacecraft was equipped with an inertial platform and digital computer to aid in orbital navigation and to provide for reentry navigation and guidance. Rendezvous radar and docking equipment were installed in the nose of the Gemini capsule. The hypergolic propellants used in the Titan II launch vehicle produced a low-yield explosion in the event of a launch vehicle breakup, so ejection seats could be used instead of an escape rocket. Finally, to accommodate extravehicular activity, a hatch operable while in orbit was provided over each astronaut's position.

In contrast to the great number of unmanned flights that preceded manned operations in Project Mercury, there were only two unmanned flights in the Gemini Program. The purpose of these flights was primarily to check out system operations and the compatibility between spacecraft and launch vehicle.

The Gemini Program also included a target vehicle to accommodate rendezvous operations. This target vehicle was a converted Agena propulsion stage that was launched into a rendezvous-compatible orbit by an Atlas launch vehicle. The procedure was to launch the target vehicle shortly before the Gemini was launched and then to maneuver the Gemini through a series of rendezvous maneuvers until the Gemini was brought within a few feet of the target vehicle. After inspection, the Gemini was maneuvered into a docking position, and the docking maneuver was made. After docking was completed, the Agena propulsion system could then be controlled by the crew on board the Gemini. The Agena propulsion stage could be used to make several maneuvers while the Gemini was attached, for instance, raising the orbit's apogee to much higher altitudes than could otherwise be achieved. There were 10 manned flights of Gemini during a 20-month period. All of these flights were successfully completed, and a great deal was learned during the Gemini Program that was later used in the Apollo Program, particularly for extravehicular activity, rendezvous techniques, mission control procedures, reentry control, and postlanding recovery operations. Figure 2 shows a picture of the Gemini spacecraft along with the adapter.

Apollo Program

Although Project Mercury was still in the early developmental phase, NASA started considering more advanced manned missions during the spring and summer of 1959. The most obvious and appealing prospect was for some type of lunar mission. Three types of manned missions were considered. They were, in order of increasing difficulty, circumlunar flight, lunar orbit, and lunar landing. Regardless of the ultimate goal, it was generally felt that the first flight would be circumlunar, and the spacecraft would pass within several hundred kilometers of the lunar far side. After this, orbital flights would be made using the same outbound and homeward navigational techniques proven in the circumlunar missions. Finally, a lunar landing would be made by descending from lunar orbit. By this scheme, each mission would be an extension of the previous one; thus the overall difficulty of achieving the final goal would be divided into a number of

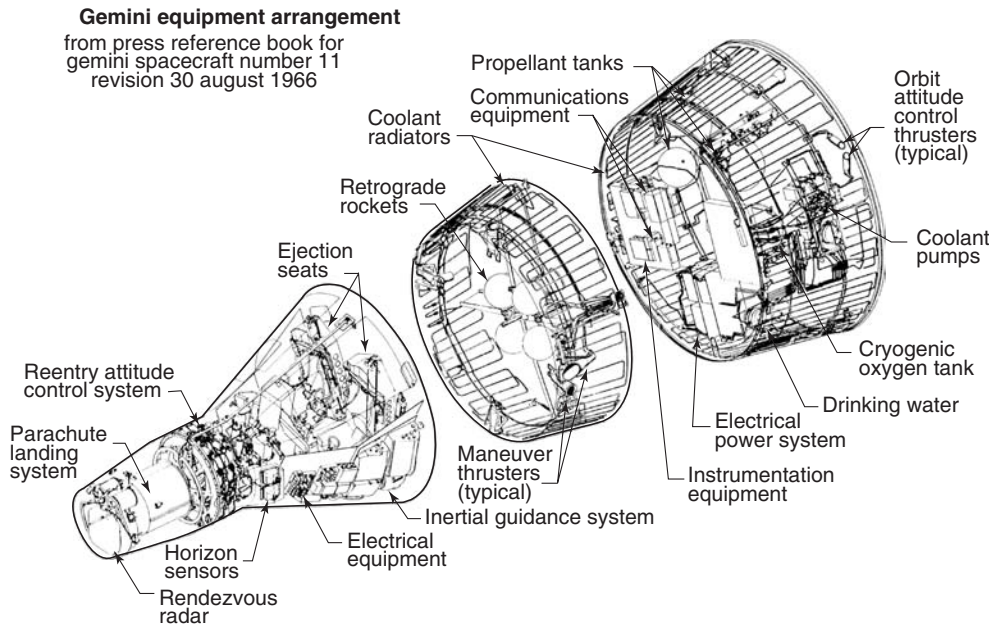


Figure 2. A cutaway picture of the Gemini spacecraft with the adapter section that attaches the spacecraft to the Titan launch vehicle. The position of the two astronauts in the spacecraft is also shown. This figure can also be seen at the following website: <http://www.hq.nasa.gov/office/pao/History/diagrams/gemini4.gif>.

incremental steps; each would have greatly reduced exposure to the unknown. Nevertheless, the first time a plan to make a manned landing by descending from lunar orbit was outlined to NASA management, several in the audience severely questioned the wisdom of not taking advantage of the experience that would be obtained from Surveyor, which was designed to go directly from Earth straight down to lunar surface. Clearly, they had not considered the excitement of the crew during a landing that started at hyperbolic velocity in a near-vertical approach and that would be fully committed before they knew whether the landing rocket would fire. This incident is mentioned to illustrate that at the same time a manned lunar landing was seriously being debated, the basic understanding of the venture was quite primitive (3).

Because it required the least amount of propulsion and the least sophistication in navigational and guidance equipment on board the spacecraft and it clearly seemed the least hazardous, the first mission seriously considered was circumlunar navigation and return. This was a modest extension of orbital flight; in fact, circumlunar flight is achievable by a highly eccentric Earth orbit of the proper parameters. However, the gravity field of the Moon creates a major influence on such orbits. Consequently, even the smallest error in the state vector at the time of translunar injection could not go uncorrected for safe entry into Earth's atmosphere at the end of the mission. The question was how to determine the error and how accurately the corrections could be made. There was real concern that a safe entry into Earth's atmosphere at lunar-return velocity might

be beyond the guidance and navigational “state-of-the-art” technology. At these velocities, the spacecraft must pull negative lift to skim along a very narrow corridor of the upper layer of Earth’s atmosphere during the initial stages of entry. If the upper boundary of this corridor were exceeded, the spacecraft would skip out of the atmosphere back into a highly eccentric orbit and perhaps expend its supplies before entering the atmosphere a second time. On the other hand, if the corridor boundary were missed on the lower side, the spacecraft would exceed the heating or load limitations of its structure. An associated concern was entry and landing point location. Because the mission was not very well understood, it was conjectured that the time of return might vary greatly from the planned time and, because of Earth’s rotation, the geographical position of the entry might have a large dispersion. For these reasons, configurations with a fairly high lift-to-drag ratio appeared desirable. In summary, the thinking in 1959 was that from the standpoint of flight control, the circumlunar mission would be flown using ground-based navigation obtained from tracking data. Guidance instructions would be transmitted to the crew for the necessary midcourse corrections. A budget for velocity changes of 152 m/s (500 ft/s), it was initially estimated, needed a lift-to-drag ratio of more than 1 to provide a large maneuvering footprint while safely staying within the entry corridors.

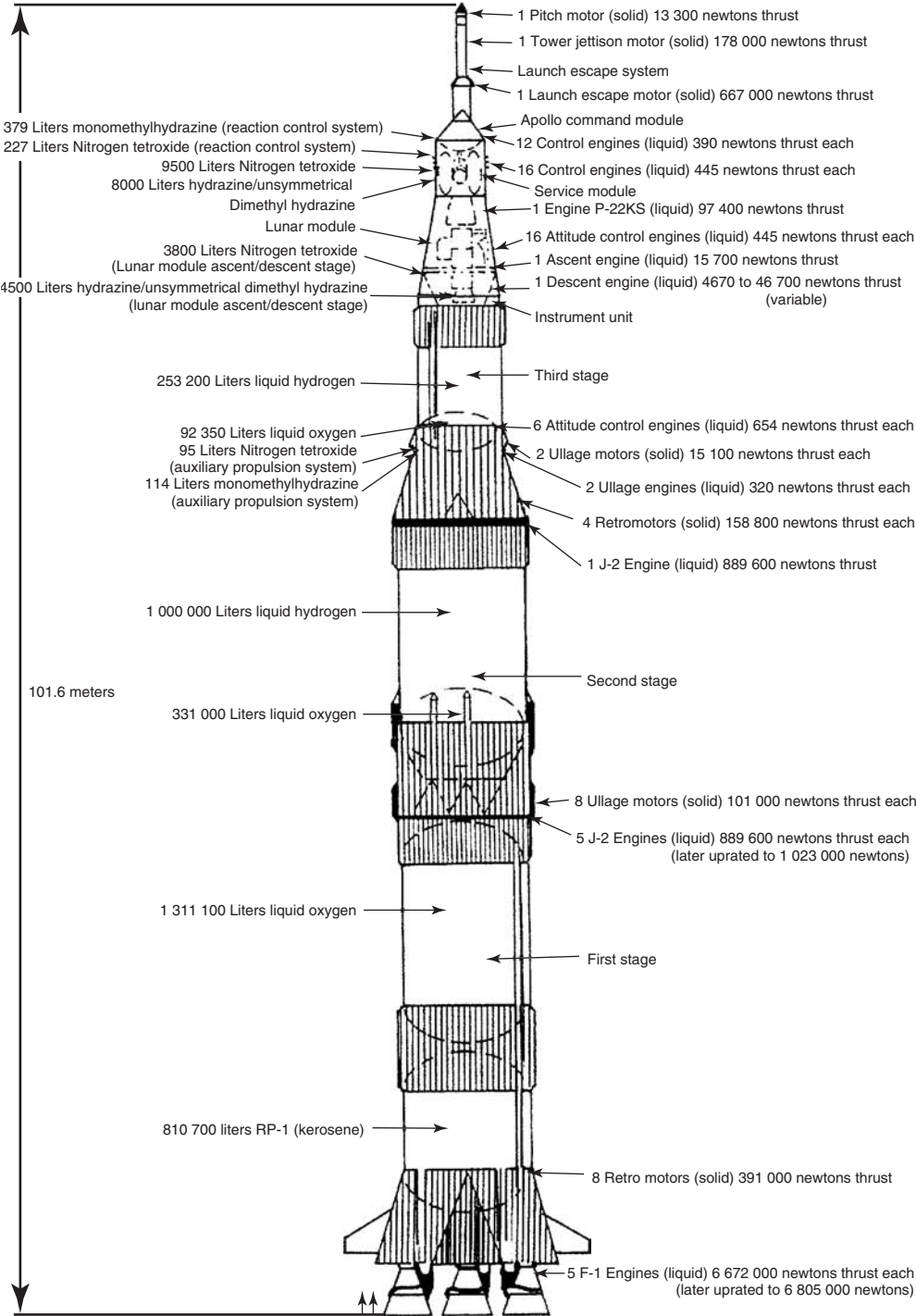
During the 1960s, enough studies had been performed by NASA and industry to understand fairly well the implications of various manned lunar mission models. Missions that involved putting a spacecraft in orbit around the Moon for a total flight duration of about 2 weeks received a great deal of interest. Such missions did not appear to be a great deal more difficult than circumlunar flight but would provide much more scientific data. Furthermore, these missions would provide the means for gaining significant flight operational experience and reconnaissance data that would support a future lunar landing. At this point, the most important roadblock to executing of a lunar mission was that an enormous new space launch vehicle would be required. Another consideration was that features of the lunar surface were not well enough known to make detailed planning of a lunar landing possible. Both the United States and the Soviet Union were conducting robotic orbiting missions that would yield high-resolution photographs to make landing site selection possible. Also, hard landers (Ranger) and soft landers (Surveyor), both U.S. missions, would reveal the properties of the lunar surface in structure and hardness.

On 25 May 1961, President Kennedy, in an address to the U.S. Congress, said, “... I believe that this Nation should commit itself to achieving the goal, before this decade is out, of landing a man on the Moon and returning him safely to Earth” (4). This commitment initiated the Apollo Program and precipitated some firm decisions. First and foremost, the return to Earth had to be determined because the safety of the crew ultimately depended upon this. The guidance and control precision of entry into the atmosphere at lunar-return velocity was sufficiently well established to commit to an entry configuration that would have a lift-to-drag ratio of 0.5 instead of the value of 1 previously mentioned. The selected value was compatible with the use of a semiballistic entry configuration design. Such configurations could achieve the relatively low entry heating of high-drag ballistic designs with a modest amount of lift. Furthermore, these features could be embodied in an axisymmetric shape, which simplified a number

of design, manufacturing, and test considerations. The design chosen for Apollo was a derivative of the Mercury shape (5). By offsetting the center of mass by a distance of 19 cm (7.5 in) from the centerline, the Apollo spacecraft would trim at about a 33° angle of attack, which was sufficient to produce the desired L/D of 0.5. Using this much lift, the spacecraft could be confidently guided to land 9360 km (5000 nautical miles) downrange from the entry point. As various equipment items were designed for the entry capsule, the center of mass inexorably moved toward the center of volume. Consequently, the center of mass ended up displaced only a little more than 12.7 cm (5 in) from the centerline and the resulting lift-to-drag ratio was lowered to 0.35. However, this turned out to be more than sufficient. Planned splashdown for all missions actually flown was set for 2593 km (1400 nautical miles) downrange, and the never used capability either decreased it to 1482 km (800 nautical miles) or increased it to 4074 km (2200 nautical miles). It should be mentioned that only once in all of the returns from the Moon was it deemed desirable to move the preplanned landing point. It was moved 926 km (500 nautical miles) further downrange to avoid the possibility of predicted bad weather at the previously chosen landing point. However, the decision to relocate was made early, and the change was accomplished by a propulsion maneuver during trans-Earth coast a day before entry. Thus, the actual entry was standard at a nominal downrange distance.

About a year after President Kennedy gave the go-ahead for the Apollo Program, a decision was reached on which launch vehicle to use. Originally, a very large launch vehicle called Nova was in the plan; Nova was substantially larger than the Saturn V rocket system that was actually employed (6). Nova was necessary because, originally, the entire spacecraft would be put down on the Moon. The judgment was made that the development of Nova would be too costly and would take long enough that President Kennedy's deadline for the lunar landing could not be met. Using the less capable Saturn launch vehicle, two spacecraft would be necessary because a rendezvous maneuver would be employed to reduce the weight of the system that would have to be lifted to the Moon (Fig. 3).

A debate ensued in 1962 whether the rendezvous between the two spacecraft should take place in Earth orbit or in an orbit around the Moon. The advantage of the former was that a spacecraft would remain in Earth orbit following the Apollo missions as a legacy, and many thought that was important. The principal advocate of the lunar orbit rendezvous was Dr. John Houbolt of the NASA-Langley Research Center (7). Eventually he made a compelling argument that only by adopting the lunar orbit rendezvous method would it be possible to meet President Kennedy's deadline of 1970 for the landing on the Moon. From the standpoints of mission planning and guidance and navigation, lunar orbit rendezvous was completely compatible with all of the work that had been done by the time the decision was taken. The requirement of the rendezvous in lunar orbit did have a major impact on the equipment and the operational techniques that would be necessary for rendezvous navigation. The experience gained during the Gemini Program was extremely valuable in that connection. Furthermore, the general rules of the interplay between the mission control center in Houston and the astronauts were also developed during the Gemini Program.



The two spacecraft required for the lunar orbit rendezvous were the command and service module (CSM) for flight to lunar orbit and return and the lunar module (LM) for descent to the lunar surface and return to lunar orbit (Fig. 4). Recognizing that the duration of the Apollo mission would necessarily be extended, a crew size of three was chosen, partly to accommodate the 4-hour-on, 8-hour-off duty cycle wherein one crew member would be on duty at all times. However, based on Gemini experience, the crew stated a decided preference to sleep, eat, and be on duty at the same time. Accommodating this preference to the mission was not considered unsafe. Ground controllers could easily monitor the condition of the spacecraft systems during the crew off-duty periods. Nevertheless, the choice of a three-man crew was well suited to the purposes of the program, and, as a matter of fact, it would have been extremely difficult to have performed the necessary tasks with less than three. Based on this number, one crewman was left in lunar orbit aboard the CSM while the other two members descended to the lunar surface in the LM. It was deemed highly desirable to have two crewmen on the lunar surface, particularly during the extravehicular activity, because this accommodated the buddy system, which materially added to the safety of that operation.

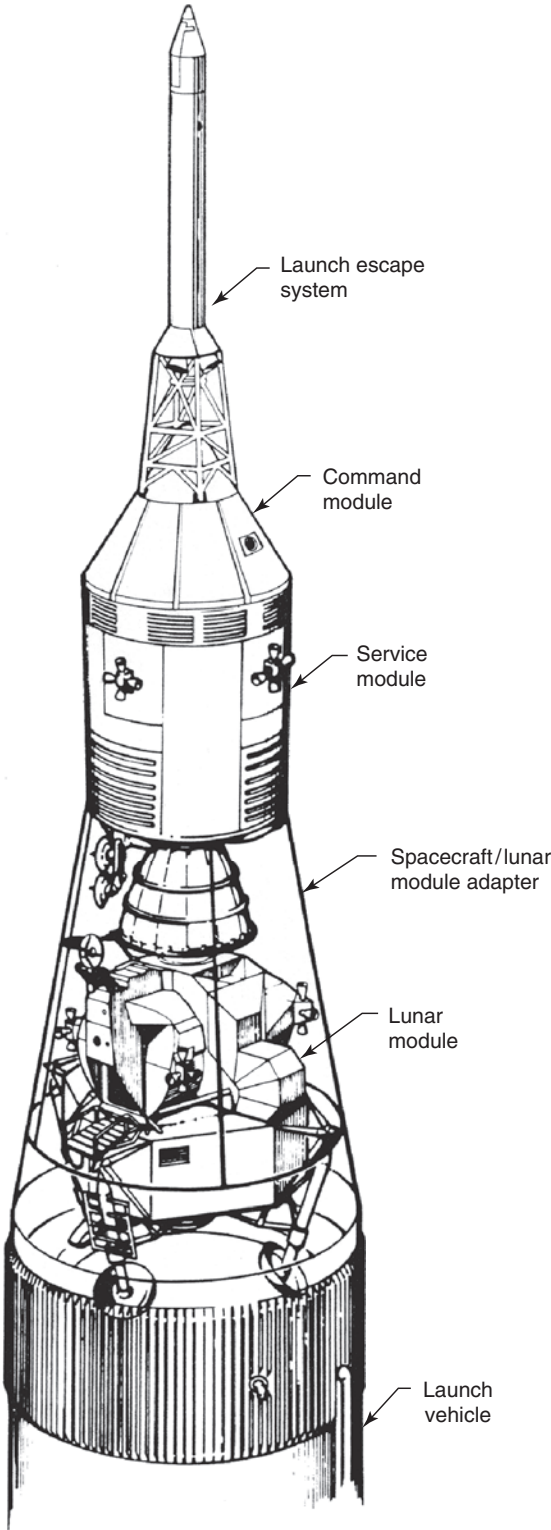
The arrangement of the CSM was quite similar to that of the Gemini spacecraft. The command module (CM) housed the crew and was the only vehicle designed to reenter the atmosphere. Like Mercury, it was equipped with a launch escape rocket to be used in case of a mishap. Although the configuration was similar to that of Mercury, the conical afterbody was blunter to minimize heating on this portion at the 33° design angle of attack. A docking mechanism and tunnel were located at the apex of the cone to accommodate docking with the LM and crew transfer. The command module was much roomier and considerably more comfortable than the Gemini vehicle. (During other periods of flight, the center couch was removed to increase room for crew activity.) There was sufficient room under the remaining couches for two crew members to sleep.

The service module, like the Gemini adapter, provided for propulsion and electric power during the mission. It was equipped with 16 reaction control thrusters arranged in four identical modules. These provided the thrust for rotational and minor translational maneuvers. The service propulsion system was equivalent to a propulsion stage. It was used during the mission for several minor maneuvers and two major ones: insertion into lunar orbit and departure from lunar orbit into trans-Earth flight. The service module also contained a thermal radiator and a high gain S-band dish antenna (Fig. 5).

The LM (Fig. 6) was really a two-stage vehicle. The descent stage was equipped with a throttleable descent engine, a landing radar, and a four-legged



Figure 3. The Saturn V launch vehicle that was used during the Apollo program to take people to the Moon. There were two stages to take the spacecraft that would go to the Moon into near Earth orbit. A third stage, the Saturn IV B, propelled the spacecraft, consisting of the command module, the service module, and the lunar module, into a lunar trajectory. The Saturn IV B was jettisoned following the trajectory insertion, and it eventually crashed into the Moon. This figure also can be seen at the following website: http://www.cr.nps.gov/history/online_books/butowsky4/images/space12c.jpg.



Apollo launch configuration for lunar landing mission

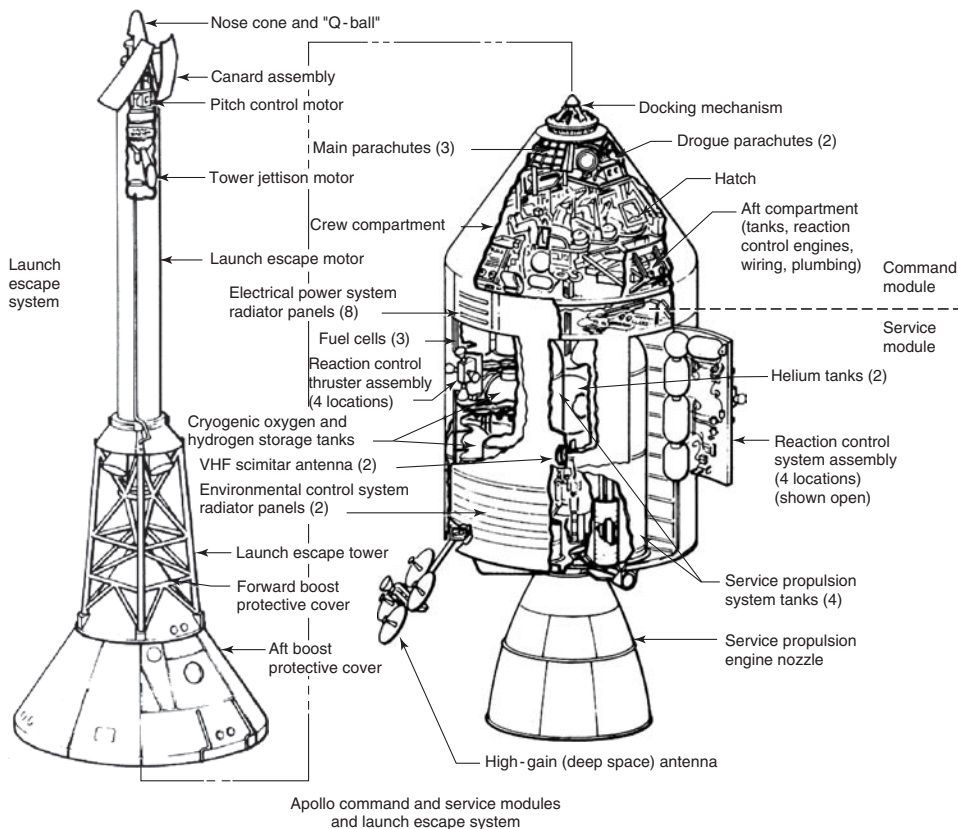


Figure 5. A detailed cut-away drawing of the Apollo Command and Service Modules. The positions of the three astronauts in the Command Module are indicated. The launch escape system is also shown. This figure also can be seen at the following NASA website: <http://www.hq.nasa.gov/office/pao/History/diagrams/ad004.gif>.

landing gear. The legs had an extra large spread to accommodate the low impact stability associated with the very low gravity of the Moon. Each leg was tipped with a large dish-shaped pad, which was designed to accommodate the unknown bearing pressure of the lunar soil. The pad was socket-mounted to allow skidding at touchdown in any direction with little chance of tripping. The descent stage was designed to act as the launch platform for the ascent stage at the time of departure from the lunar surface.

The ascent stage housed the ascent propulsion system and the LM cabin. There were accommodations for two crewmen, who were positioned at their flight



Figure 4. A picture of the Apollo spacecraft in the launch configuration. The Command Module, the Service Module and the Lunar Module are clearly illustrated. The legs of the Lunar Module folded so that the vehicle fit into the conical shroud on top of the Saturn V spacecraft. The escape tower is also shown. This figure also can be seen at the following NASA website: <http://www.hq.nasa.gov/office/pao/History/diagrams/ad003.gif>.

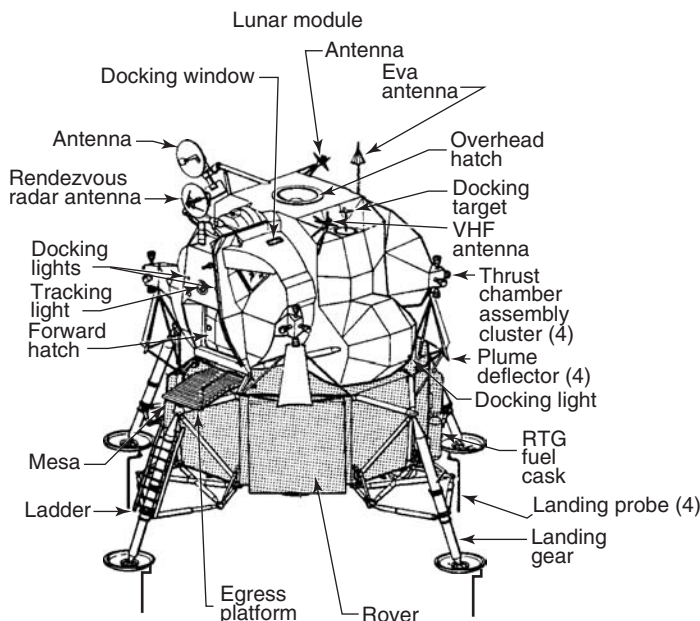


Figure 6. A drawing of the Lunar Module in the landing configuration. The shaded portion is the descent stage which contains the rocket motor used for the landing. This portion of the spacecraft is left on the lunar surface following the completion of the extravehicular activity (EVA). The ascent stage also has a propulsion system that moves this vehicle back to lunar orbit where it then mates with the Command and Service Module. This figure also can be seen at the following NASA website: <http://www.hq.nasa.gov/office/pao/History/diagrams/ad014.gif>.

stations in the standing position to maximize downward visibility during the landing maneuver. The crewmen slept in hammocks while on the lunar surface. Like the Mercury and Gemini spacecraft, the CSM and the LM had a pure-oxygen atmosphere at a pressure of one-third sea-level pressure. The ascent stage was equipped with a rendezvous radar, 16 reaction control thrusters, and a high-grain S-band dish antenna. The LM was battery-powered and water-cooled. Ascent propulsion was used to launch the ascent stage from the lunar surface into a rendezvous orbit. It was also available to abort the landing maneuver into a rendezvous orbit if the landing could not have been made.

After the CSM had established the desired orbit about the Moon, the LM pilot and the commander transferred to the LM and powered it up. The LM was then separated and the CM pilot was left alone in the CSM until the other crewmen returned from the lunar surface. The LM then made a phasing maneuver, by which its orbit relative to the CSM was lowered. This maneuver placed the LM sufficiently ahead of the CSM so that an aborted powered descent would terminate into a rendezvous-compatible orbit. At the proper time, powered descent was initiated with the descent propulsion system at nearly full throttle. Slight throttle adjustments were made during descent to keep the descent

trajectory on target. As the targeted landing area was approached, the LM was pitched over while, at the same time, the engine was throttled back. This allowed the crew to achieve visual contact with the landing area several minutes before touchdown. At this time, the commander could determine the touchdown point that was being targeted by the onboard guidance system. He did this by looking through a scale painted on his window, called the landing point designator. He could then override the guidance system until the landing point designator indicated a desirable landing location. During this period, altitude and descent rate were sensed by the landing radar, and control by the guidance system was continued. Hovering flight was achieved several hundred feet above the lunar surface. The final descent and touchdown were then made under control of the commander. As the LM approached the surface, a considerable amount of dust was blown up, but some visibility remained, and good landings were made in all cases. It should be mentioned that several turbofan-powered hovering vehicles designed to mimic the landing characteristics of the LM were built. The crew used the lunar landing training vehicles extensively to sharpen their proficiency before each lunar landing mission.

While on the surface of the Moon, the crew used extravehicular mobility units (EMUs) to move about. The EMUs each consisted of a space suit pressurized to 25 kPa (0.25 atm) and a battery-powered backpack. In many ways, the EMUs functioned like miniature manned spacecraft. They provided a cooled, revitalized atmosphere for the astronaut to breathe. The astronaut was kept at a comfortable temperature level by an actively cooled garment through which water was circulated at a regulated temperature. The EMU was equipped with redundant two-way communication links that provided voice communication between the two astronauts and, using the LM communication equipment as a relay station, communication with the Mission Control Center on Earth. Ground controllers could also monitor the physical condition of the astronauts and the performance of the equipment by telemetry from the EMU.

The suits provided sufficient mobility to accommodate a range of activity on the lunar surface such as setting up science stations, taking photographs, and performing short geological traverses for sample selection. On later missions, lunar surface exploration was greatly enhanced by the Rover, an electric-powered four-wheel vehicle that could carry the astronauts and a considerable amount of equipment. The Rover carried two-way communications equipment and a high-gain antenna that provided sufficient bandwidth for a direct link to Earth carrying color or television transmissions as well as relaying the communications from the astronauts to Earth.

It was recognized that both the CSM must be made considerably more reliable than previous spacecraft for several reasons. Compared to an orbital flight, a journey to the Moon implied a much greater exposure to the hazards of space. As a consequence of the complexity and the time required to abort a lunar mission, both the LM and the CSM had backup navigational and guidance systems, adequate reserve propellant, and levels of redundancy that were carefully determined to be sufficient. All possible failure modes were analyzed so that single-point failures were identified and eliminated when possible or safely accommodated. Furthermore, operational methods were worked out to circumvent failures or to ameliorate the conditions that might be brought about by failures.

Although tracking from the ground was chosen as a primary method of navigation during the mission, it was decided that the spacecraft should be capable of autonomous navigation and guidance to return safely if all communication links were lost. The onboard navigation and guidance system was also used in verifying the accuracy of various maneuvers. Navigation in space, as on the ocean, requires precise determination of position. For this purpose, the Apollo command module was equipped with a sextant to measure the angle between a number of preselected landmarks on both Earth and the Moon and certain cataloged celestial bodies. Sightings and the time of sightings were fed into the onboard computer, which was preprogrammed to solve the navigational problem. The spacecraft was also equipped with inertial measurement units, which, together with the computer, performed the automatic guidance function.

In addition to the launch vehicle and LM and CSM, there were two other vital elements in successful execution of the lunar missions. These were the Manned Space Flight Network (MSFN) and the Mission Control Center, including all of the mission procedures and time lines that were worked out by Mission Control Center personnel. The MSFN was modeled after the Deep Space Information Facility (DSIF) developed by the Jet Propulsion Laboratory. The MSFN consisted of three 25.9-m (85-ft) diameter antennas located at approximately 120° intervals of longitude around Earth to provide coverage of the mission.

Both the MSFN and the hardware onboard Apollo were capable of producing highly accurate navigational data. Data from both sources were compared before making any velocity change maneuver. Also, immediately after the maneuver was made, data were again cross-checked. Navigation done on the ground had the benefit of a large complex of powerful computers. Furthermore, at least two S-band trackers were always available as data sources. The data from the S-band tracker were extremely accurate. In addition to providing a Doppler count for velocity, the carrier signal was also phase-modulated with a pseudo-random noise (RN) code for range measurement. This digital signal, which was nonrepetitive for 5.5 s, was turned around and retransmitted on another carrier by a transponder on the spacecraft. Distance measurements as accurate as about 10 m could be obtained. Velocity measurements were much more useful. High-powered data processing techniques could produce an accuracy better than a millimeter per second by smoothing Doppler data across a period of 1 min. Using such data, extremely accurate state vectors could be obtained not only on trans-lunar and trans-Earth flight but also while Apollo was in lunar orbit. This capability was important because lunar gravitational anomalies and venting from the spacecraft continually perturbed the orbit. Computational techniques were developed to the point at which tracking data obtained from the lunar module during its landing descent burn could be processed to serve as a sufficiently accurate "tie-breaker" if onboard primary and backup computations produced unexplainable differences.

The general approach to mission planning was to break the mission down into a number of discrete events and periods. Each of these was analyzed in great detail, and a complete model of the mission at great precision was constructed before flight. When flown, the missions would usually duplicate the plan in exact detail. A feature of the planning was the inclusion of time allowances for

unexpected events so that the preplanned schedule could be maintained. The advantage was that almost every event or phase of the basic mission was extremely well understood and exercised. In addition to the basic mission plan, a great number of contingency plans was available to cover every rational problem.

All missions were planned to accommodate midcourse corrections both outbound and on return. Specific times were set aside for these maneuvers: four translunar midcourse correction events were allowed and three trans-Earth. However, if the error to be corrected was sufficiently small, the maneuver would not be made, and, as a matter of fact, many missions needed only one corrective maneuver each way. When first considering translunar flight in 1959, the budget of 142.4 m/s (500 ft/s) for midcourse corrections was established. The estimate was down to 91.4 m/s (300 ft/s) when the program was initiated several years later. When actual flight began, the "three sigma" estimate was 23.8 m/s (78 ft/s). Actually, most flights required less than 6.1 m/s (2.1 ft/s). For example, on the last flight, Apollo 17 executed only one correction maneuver each way: translunar, it was 3.2 m/s (10.6 ft/s); for trans-Earth, only 0.6 m/s (2.1 ft/s) was needed, which shows the great improvement in navigational systems.

The robotic missions that preceded the Apollo landings, Rover, Surveyor, and Lunar Orbiter have already been mentioned. The Ranger spacecraft was a probe that transmitted a few close-up images of the lunar surface just before colliding with the Moon at high velocity. The Surveyor was a soft lander that made five successful landings on the lunar surface. After landing, the Surveyor transmitted pictures of the surrounding features of the moonscape that provided extremely useful information on surface roughness as well as the quantity and size of rocks. Just as important, engineering data obtained from the Surveyor landings were extremely valuable in verifying the firmness of the lunar surface for landing the lunar module. The unmanned Lunar Orbiter flights, however, were every bit as valuable to the Apollo Program. They provided high-resolution photographs of the lunar surface that were extremely useful in selecting landing sites. The cartographic quality of the photographs was more than sufficient to make accurate maps of the lunar surface that could be used for orbital navigation and for visual recognition by the astronauts in the terminal phase of their descent. Just as important, analysis of orbital tracking data greatly improved the accuracy of the lunar gravitational constant and provided valuable data on lunar gravitational anomalies; all of them facilitated translunar and lunar-orbit navigation on the Apollo missions. The accuracy of navigational techniques that were ultimately developed made it possible for Apollo 12, the second landing mission, to come to rest within a short walking distance of Surveyor II, which had landed on the Moon $2\frac{1}{2}$ years previously. The Surveyor's location had been identified on a Lunar Orbiter photograph by patient and meticulous study.

Twelve unmanned test flights were made in the Apollo development program. All but one was successful. Six of these were devoted to qualifying the launch escape and parachute deployment system for the command module. The others were concerned with systems tests of the spacecraft and with compatibility of the launch vehicle and the launch environment. In one of these tests, the service propulsion system was used to drive the command module back into the atmosphere at the velocity and in the flight path expected during lunar return.

Eleven manned flights were made in the Apollo lunar program. Nine of them were journeys to the Moon, and six were lunar landing missions. These flights are summarized in Appendix B.

On 27 January 1967, the program underwent a critical setback during a countdown rehearsal for what was planned to be the first manned Apollo flight. The interior of the command module suddenly burst into flames, and trapped the crew. Astronauts Gus Grissom, Ed White, and Roger Chaffee were killed, and the spacecraft was destroyed. Although the immediate source of the fire was never determined, the general cause was associated with the use of many materials in the cabin interior that were flammable in the pure-oxygen cabin atmosphere, which, during prelaunch conditions, was at sea-level pressure. It took the program more than a year to recover from the fire (8).

Extensive changes in materials were made, and in cases for which suitable replacement materials could not be found, fireproof coatings and coverings were used. Special test programs were made to certify the fireproofing program. It was found that, although good fire protection was achieved for an oxygen atmosphere at the reduced pressure of orbital flight, there were no practical solutions for the prelaunch conditions when the cabin would be at sea-level pressure. Therefore, the oxygen in the cabin was diluted with 40% nitrogen during prelaunch activities. After the cabin pressure dropped to one-third of sea-level pressure during ascent, there was still sufficient oxygen for breathing. Later, when the astronauts were ready to depressurize further to spacesuit pressure, the nitrogen content of the atmosphere had been sufficiently purged to preclude a bends problem.

When astronauts James Lovell, Jack Swigert, and Fred Haise were on board Apollo 13, which was to have been the third lunar landing mission, another accident occurred. During translunar flight as the Moon was being approached, one of the two tanks used to store cryogenic oxygen in the service module failed. The failure was sufficiently violent to cause the oxygen in the other bottle to begin leaking, most probably through an external line. Except for gaseous oxygen stored in the command module for life support during the short period of reentry flight, there was no other oxygen supply for the CSM. The lost oxygen was to be used both for life support and to power the fuel cells. The failure was eventually traced to a thermostat switch that had been frozen shut because of an overcurrent applied during the preflight checkout on the launch pad. As a result, the heater wire inside the oxygen tank became too hot, and the fluorocarbon insulation melted causing a spark. This ignited the insulation which readily burned in the pressurized oxygen tank, and the resulting pressure rise caused the tank to fail.

This desperate situation was countered by moving the crew to the LM cabin and powering down the CSM. Because the LM was battery-powered for only a planned 3-day period, it was configured for minimum power consumption. After getting everything in order, the crew made a velocity change maneuver to correct the flight path to a "free-return" trajectory by which the spacecraft would swing around the Moon and head back to Earth on a path consistent with reentry targeting. Another major propulsive maneuver was made after the spacecraft was in trans-Earth flight. That maneuver changed the trajectory to shorten the return time. Although there was sufficient oxygen for the return flight, an improvised method for adapting the command module's lithium hydroxide canisters

used to remove carbon dioxide for use in the LM environmental control system had to be devised. This was done by experimenting on the ground and then communicating instructions to the crew. Fortunately, the crippled spacecraft returned safely with an unharmed but disappointed crew. The Apollo program continued for four additional and successful lunar landings.

The Skylab Program

After the completion of Apollo 17, the Apollo CSM and the Saturn launch vehicles were used for the Skylab Program (Fig. 7). The third-stage structure of the Saturn V was converted to a space laboratory. The hydrogen tank of the Saturn IV B stage was divided into laboratory, sleeping, eating, recreation, and hygiene compartments. Storage lockers, an environmental control system, and other equipment were added for the comfort and physical well-being of the crew. The oxygen tank was converted to an oversize trash container. A deployable micro-meteoroid shield and passive thermal control coatings were added to the skin of the tank. A battery of telescopes designed to study the Sun were mounted on a high-precision pointing platform. There was also an airlock to accommodate EVA and a docking module to which the CSM could dock. In addition to the solar telescopes, a great variety of experimental equipment was carried, including a number of Earth-pointed remote-sensing instruments and cameras.

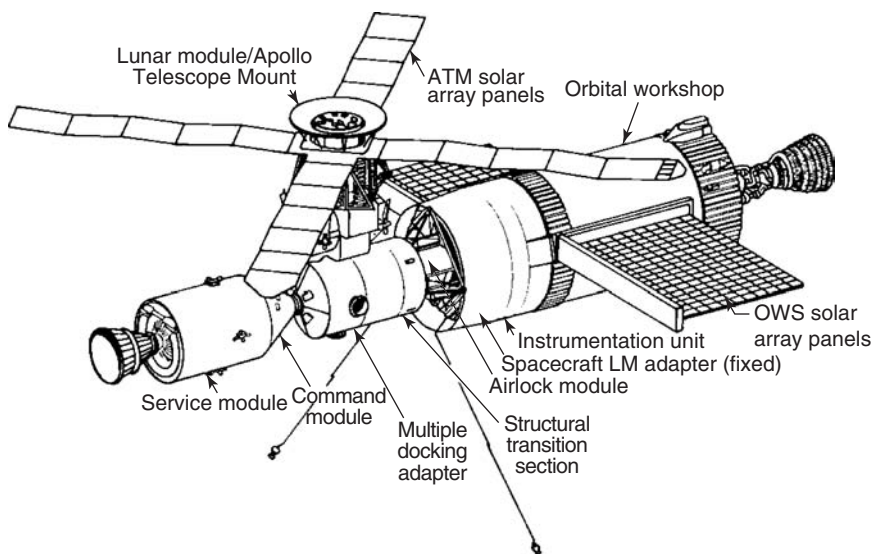


Figure 7. The Skylab orbital workshop configuration. The Apollo Command and Service Module is on the left, the center section is the Apollo Solar Telescope Mount, and the left-hand section shows the Modified Saturn IV B that became the living space for the three astronauts during the three missions when they worked in this temporary space station. This figure also can be seen at the following NASA website: <http://www.hq.nasa.gov/office/pao/History/SP-4011/p144.htm>.

The Skylab orbital workshop was stabilized using three control moment gyros. These could be desaturated from time to time by reaction control jets using nitrogen as a propellant. Electric power for the Skylab was provided by two photovoltaic solar-powered systems. One system consisting of four folding arrays was attached to the solar telescope mount. The other system consisted of two wing arrays extending from the converted fuel tank (10).

The Skylab orbital workshop was launched on 14 May 1973 into a 50° inclination orbit. The nearly circular orbit ranged in altitude from 496 to 498 km (268 to 269 nautical miles). During launch, one of the wing solar cell arrays and the micrometeoroid shield were carried away. The other wing was jammed so that it could be only partly deployed. The micrometeoroid shield also was important in the passive thermal control system; its absence left the tank skin exposed. This was coated with a highly reflective coating which happened to be biased hot. The result was that the laboratory quickly became overheated and reached internal temperatures higher than 50°C.

The launch of the first crew was delayed while special repair equipment was designed, constructed, and tested on the ground. Eleven days following the launch of the Skylab orbital workshop, astronauts Pete Conrad, Joe Kerwin, and Paul Weitz were launched on Skylab 2. They deployed a parasol-like sunshield over the exposed skin which reduced the interior temperature to 24°C. They also freed the jammed wing of the solar cell array. Although the missing wing reduced the power available from that expected, there was sufficient power to perform the planned activity.

The Skylab orbital workshop was occupied by three different crews (See article on Skylab elsewhere in this Encyclopedia). Each successive crew stayed a longer period of time. A great number of experiments and science investigations was performed before the Skylab orbital workshop was abandoned on 8 February 1974. It was left in a condition that would allow partial reactivation in a possible visit from the Space Shuttle. Unfortunately, its orbit decayed before that was possible. Skylab reentered the atmosphere on 11 July 1979, and came to Earth mostly in the Indian Ocean, with some pieces landed in western Australia.

Apollo-Soyuz Project

After a series of meetings in 1971 and 1972, the United States and the Soviet Union agreed to a joint manned space mission as part of the “détente” policy pursued by the Nixon administration. This mission became known as the Apollo-Soyuz Test Project (ASTP) (Fig. 8). Its basic purpose was to produce hardware and operational procedures that would provide the means for one country to work with the other in future manned space missions. Particular emphasis was placed on assistance and rescue missions. Rather than merely trade design and mission control information, it was agreed that the only way to be sure assistance and rescue missions would really be feasible would be to execute an actual joint mission (11).

It was agreed that the American Apollo and the Russian Soyuz would rendezvous and dock during a cooperative mission. To accomplish this objective, a number of technical problems had to be addressed. The primary problems were

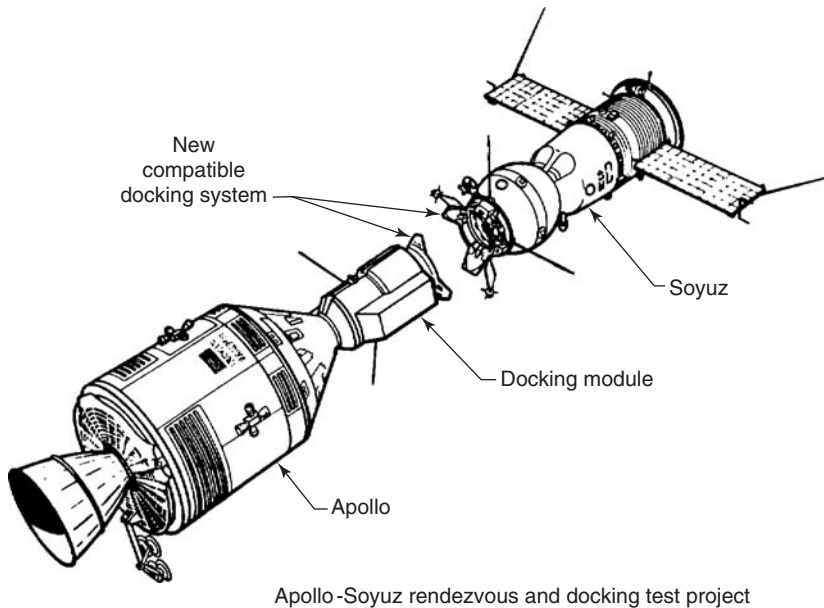


Figure 8. The Apollo-Soyuz spacecraft. The Apollo Command and Service Module, the Docking Module, and the Soyuz capsule are all clearly shown. Two Soviet cosmonauts and three American astronauts performed the docking maneuver that led to the first meeting in space of people from each nation. Experiments and tests performed by the crew led to naming this effort the Apollo-Soyuz Test Program. This figure can also be seen at the following NASA website: <http://www.hq.nasa.gov/office/pao/History/diagrams/astp/pk69.htm>.

compatible communications both in space and on the ground, compatible mission control procedures, compatible docking hardware, and accommodations for the different atmospheric constituents in the two spacecraft. The Soyuz cabin was maintained at sea-level pressure with the same 80%/20% mix of nitrogen and oxygen found in air. The Apollo cabin was at one-third sea-level pressure using an atmosphere of 100% oxygen. This incompatibility was solved by using a docking module that would be carried with the CSM during launch. It was carried in the adapter in a manner similar to that used to carry the LM during lunar missions. On one end of the docking module was the international docking system, which had been agreed on with the Soviet Union. The other end, like the LM, had the portion of the Apollo probe and drogue docking system. The docking module had an interval volume sufficient for as many as three people at the same time. It was also equipped with bottled supplies of both nitrogen and oxygen. Consequently, it could serve as an airlock between the two spacecraft.

The ATSP mission was successfully completed in July 1975 by the launch of cosmonauts Alexei Leonov and Valery Kubasov in Soyuz followed by astronauts John Stafford, Vance Brand, and Deke Slayton in Apollo. After rendezvous, both the American and the Russian mechanisms performed successful dockings. After the final undocking, both crews spent additional time in space performing individual experiments.

The Space Shuttle Program

The Space Shuttle system had been developed to improve accessibility greatly to space and thereby to facilitate many opportunities for space applications that would not be possible otherwise (12). Through improved operational capability and flexibility, the Shuttle would bring about a wide variety of missions and activity in space that previously had not been considered. The Shuttle missions primarily fall into three general categories:

1. Missions lasting from one to several weeks and employing one or more of the Spacelab modules developed by the European Space Agency. These missions consist of a variety of experiments and observations requiring extensive participation by the crew.
2. Missions employing one or more additional propulsion stages to carry spacecraft to orbits beyond the performance capabilities of the Shuttle alone. Most of these missions are aimed at geosynchronous orbit; some planetary exploration and military missions were also launched by the Shuttle.
3. Missions in which the Shuttle deploys satellites directly into orbit. The Shuttle can retrieve or service some of these satellites, thereby adding a new dimension of utility. In some cases, the satellites will be equipped with modest propulsive capability allowing them to move back and forth to orbits beyond the operating altitude of the Shuttle. The first example of this capability was the retrieval and repair of the "Solar Max" satellite in 1984 and later, the spectacular success of the in-orbit repair of the Hubble Space Telescope's optics in 1993. The most recent operation was the repair and refurbishment of the Hubble Space Telescope in 2002 (13).

Most Shuttle missions probably have been a combination of these general types to enable more comprehensive use of the cargo load of each flight.

When launched from Cape Kennedy, orbits with inclinations from 28.5° (the latitude of the launch site) to 56° will be achievable. The Shuttle was initially capable of carrying as much as 29,000 kg (32 tons) of payload into low-inclination, low-altitude orbits. It was capable of returning as much as 14,500 kg (16 tons) from orbit. However, various restrictions have been placed on the payload capability of the Space Shuttle since operations of the vehicle were initiated in 1981.

The Space Shuttle in the launch configuration consists of four major elements: an Orbiter, an external tank, and two solid-rocket boosters (Fig. 9). The boosters provide the majority of the thrust for the first 2 minutes of flight. Each booster weighs about 567,000 kg (1,250,000 lbm) and produces a peak thrust in excess of 11.1 MN (2,500,000 lbf). The boosters are jettisoned after burnout and are recovered for reuse after descending by parachute into the ocean near the launch site.

The external tank contains 530 cubic meters (140,000 gal) of liquid oxygen and 1438 cubic meters (380,000 gal) of liquid hydrogen. These propellants, with a combined weight in excess of 680,000 kg (1,500,000 lbm), feed the three main engines in the Orbiter. Eight minutes after liftoff, when main propulsion burnout

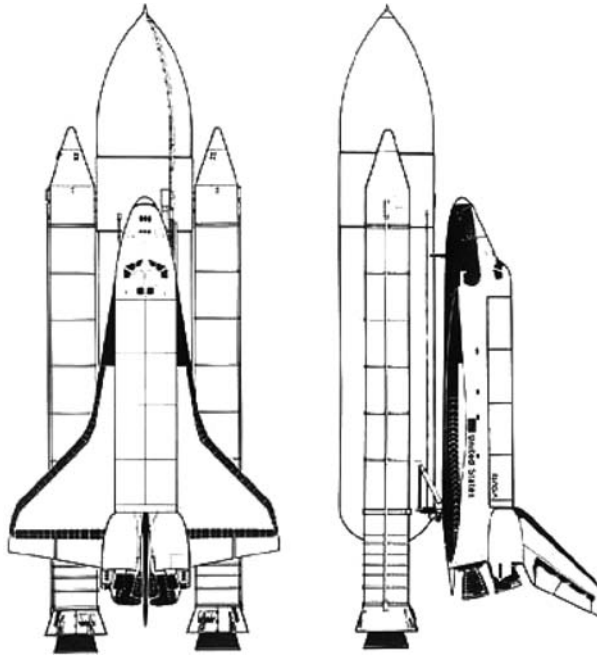


Figure 9. The configuration of the Space Shuttle system with the Orbiter, the two solid rocket motors, and the external tank for the fuel and the oxidizers for the Space Shuttle main engines mounted in the Orbiter. This figure also can be seen at the following NASA website: <http://www.hq.nasa.gov/office/pao/History/rogersrep/v1p3.jpg>.

occurs, the Shuttle is just short of attaining orbital velocity. The tank is then jettisoned and follows a long, shallow trajectory that ends in a remote portion of the Indian Ocean. The tank is destroyed during reentry. After the tank is jettisoned, the Orbiter employs the two orbital maneuvering engines to propel itself into orbit. The orbital maneuvering engines (OMES) are also used for other maneuvers in space during the remainder of the mission.

The Orbiter is the most complex part of the Shuttle and provides control of the boosters and the external tank during launch. The Orbiter is designed to be reusable; only a minimum amount of refurbishment and maintenance are required between flights. Almost all of its systems should function reliably after a hundred or more flights. Insofar as practical, development of new systems or technology was avoided in designing of the Orbiter. The Orbiter's electric power is produced from hydrogen and oxygen in fuel cells that are an improved version of those used during the Apollo Program. The basic structure of the vehicle is primarily aluminum; skins, stringers, frames, and ribs are in an arrangement typical of that of an airplane. The landing gear and the cockpit arrangement are also quite similar to those found on conventional aircraft. Hydraulic power is used to move flight control surfaces and for ground steering and braking.

The Orbiter is a key element of a highly complex launch vehicle. It is a unique and highly versatile spaceship that has astronauts on board to operate the vehicle and to perform experiments. It is the combination of these functions—

launch vehicle, spacecraft, and airplane—that make the Orbiter a most complex machine. Consequently, major advancements in the state of the art were necessary in several areas. These advances were most pronounced in flight control, thermal protection, and liquid-rocket propulsion technology. (See the article on Space Shuttle Orbiter elsewhere in this Encyclopedia.)

The external tank is 45.7 m (150 ft) long and 8.5 m (28 ft) in diameter. The structure of this huge tank weighs only slightly more than 31,750 kg (70,000 lbm). Nevertheless, the tank is the central element of the launch configuration, which has a combined weight of nearly 2,000,000 kg (4,500,000 lbm). The consequence is the creation of a number of easily excited low-frequency structural modes. Therefore, the flight control system employs sophisticated mode-suppression techniques in gimbaling the rocket engines as the vehicle is steered during launch.

Flight control during entry is even more complex. Unlike previous manned spacecraft such as Apollo, the Orbiter actually is flying during atmosphere entry and depends on active aerodynamic control surfaces to maintain a stable attitude and control of its flight path. When aerodynamic flight begins, it is flying four times faster than the X-15, the previous speed record holder.

The Orbiter's configuration features a double delta wing, a single, high, vertical fin, a huge cargo compartment, and a large aft end containing an assembly of rocket engines. This configuration is based on a compromise satisfying a large number of requirements. The cargo bay size was derived from a survey of potential traffic that clearly indicated that space cargo would in general consist of low-density objects. Large main propulsion engines were desirable to minimize the size of the boosters, which represent a major part of the cost per flight. The result was a center of mass that is unusually far aft. Depending on mass and location of cargo, the Orbiter occasionally encounters neutral to slightly negative static stability about the pitch axis at low speeds. For control during aerodynamic flight, the Orbiter is equipped with very large elevons on the wing trailing edge. A large flap extends from the rear of the fuselage to trim the vehicle. A combination rudder and speed brake on the trailing edge of the vertical fin is used for directional and speed control.

Return from orbit is initiated by making the typical retrograde propulsion maneuver using the OMES engines. The Orbiter then loses altitude and gradually enters the atmosphere. Initial control, similar to that of Apollo, is by rolling about the velocity vector at a very high angle of attack using the reaction control system. As aerodynamic forces gradually increase, the control of the vehicle becomes manageable using aerodynamic control surfaces, and, except for those controlling yaw, the reaction control jets are deactivated. The rudder, which might be expected to control yaw, is ineffective as a control surface on the Orbiter because of the high angle of attack. Both yaw jets and the rudder were active for yaw control from Mach 3.5 to Mach 1.0. Below that speed regime at a lower angle of attack, the yaw jets were deactivated, and only aerodynamic control surfaces were employed in the flight control system.

Reentry at the end of the flight is made using a preselected angle-of-attack (α) profile. An angle of attack of 40° is maintained until the Orbiter decelerates to $M = 11$. From that point on, α is decreased continually and passes through $M = 1$ at $\alpha = 8^\circ$. The approach and landing maneuvers in the subsonic flight regime

ranged in α from 4° to 10° . The acceptable reentry altitude corridor is approximately plus or minus 1500 m (5000 ft) of the nominal flight path. The Orbiter is guided to stay within this corridor, while flying on the preselected α profile, by varying the bank angle. The nominal angle at the start of reentry is 80° . This value gradually diminished to about 45° . During the first flight, the landing site was essentially straight downrange. Consequently, some portions of reentry were flown banked left and others banked right. In this case, bank-reversal maneuvers were made near numbers of $M = 18, 9, 5$, and 2.5 . In future flights, longer portions of reentry may be spent banked in one direction or the other to accommodate landing at sites which may be displaced as much as 2037 km (1100 nautical miles) on either side of the straight downrange track. Downrange navigation is achieved by flying nearer to the lower or upper bound of the corridor. This can be accomplished by using a slightly greater or less than nominal bank angle, as the case may require. During the first Shuttle flight, John Young went to control-stick steering during the last two bank-reversal maneuvers. This allowed him to enter the bank-reversal maneuver more gradually than possible when the maneuver was programmed in the autopilot. Recent simulations had shown that the stability margins could be increased by decreasing the roll-rate acceleration during this flight regime. The decision to have the crew make this maneuver avoided modifying mature software. The entry is flown by computer until subsonic flight when the pilot uses control-stick steering and makes the approach and landing.

The Orbiter configuration was subjected to more hours of wind tunnel testing than any other flying machine. This was necessary because the Orbiter flies across a greater range of Mach numbers, Reynolds numbers, dynamic pressures, and angles of attack than previous machines. More importantly, during its first flight, it was totally committed to successfully negotiating this wide and diverse range of flight conditions before it could land. This was a unique situation not previously encountered in modern aviation flight-testing. High-performance airplanes and in fact virtually all newly designed airplanes are extensively flight-tested before maximum performance is approached. Each flight test is a cautious extension of previously encountered conditions and is followed by thorough analysis of test results from which the next safe increment in the flight-test program can be defined.

As a substitute for testing in actual flight, the Orbiter's flight control system had been extensively tested in computer-driven simulation facilities. Mathematical models resident in the facility software mimic the atmospheric conditions of flight, including gusts and crosswinds. Other models represent the wind-tunnel-derived aerodynamic responses to the simulated flight environment. However, wind-tunnel data have limited precision. Results obtained in different facilities sometimes were significantly different. Furthermore, experience has shown that data obtained in flight often exceed the bounds of data scatter of wind-tunnel data. Extensive statistical analyses of these effects on the stability and the control surface effectiveness were made to determine possible worst case aerodynamic qualities. The Orbiter's flight control system successfully "flew" simulated entries in which such worst case conditions were modeled.

The heart of the Orbiter's flight control system is a set of five identical general-purpose computers. Each computer in the set has the capacity to control

the entire flight, including the guidance and navigation functions, without assistance from the others. Under control of the central computers are a great number of subordinate data processors that have specialized functions. The central computers formulate steering commands for the gimbal actuators during launch, the reaction control jets during space flight, and the yaw jets and aerodynamic surfaces during aerodynamic flight. These steering commands can originate in response to stick inputs from the crew or be generated within the central computers in the fully automated flight mode. Redundant sets of rate gyros, accelerometers, inertial measurement units, star trackers, radio navigational aids, radar altimeters, and dynamic and static pressure sensors all feed data to the central computers. The more critical sensors are quadruply redundant and allow at least two and possibly three sensors in a particular set to fail without degrading the performance of the flight control system.

The five centralized computers are divided into a redundant set of four and the fifth computer is held in reserve as an independent backup. Normally the redundant set will control the flight. Each of the four computers in this set is loaded with identical software, and all work in parallel with each other processing data from the sensors and transmitting commands to the controls. If one of the sensors should fail, each computer will detect it, and the sensor will be deactivated. Upon completion of every computation cycle, the redundant computers make a simple arithmetic comparison of their output. Any computer that fails to agree with its colleagues for two successive cycles is considered to have faulted and is deactivated. The computers also have built-in fault-detection features which should cause them to self-deactivate in most failure cases. Simple logic would indicate that this system with its depth of redundancy and its fault-detection and deactivation features should prove extremely reliable. Nevertheless, there is concern that despite all precautions, an inherent weakness could exist in the software that would ripple through the computers and deactivate them one after another. Such should an event occur, the crew would switch control to the backup computer, which is loaded with software that, although coded differently, can replace all of the functions of the redundant set.

The use of digital computers in the Orbiter for flight control provides a major improvement in versatility and precision over other systems. The digital system facilitates redundancy management, fault detection, and fault isolation and, when proven, should provide a major improvement in flight safety. Although the inherent complexity of the Shuttle flight control task forced the development of an unusually elaborate flight control system, this pioneering effort proved to be a major step toward the production of more economical and safer aircraft.

The propellant combination of liquid hydrogen and oxygen is the most energetic considered practical for use. The Shuttle, like the Saturn launch vehicle upper stages, employs these propellants for its main propulsion. However, a new high-performance engine was developed for the Shuttle (Fig. 9). It features a combustion chamber pressure of 22,409 kPa (3250 psi) at full power, or about four times higher than that of the Saturn engines. It also incorporates a new cycle that circumvents the wasteful use of some of the propellant just to power the turbine-driven propellant pumps. It might be mentioned that the total horsepower required to pump propellant to the Orbiter at full power is greater than the total propulsive power of a Forrestal-class carrier. The Shuttle rocket engines

produce 7% more propulsive energy per pound of propellant than the hydrogen-fueled rocket engines used on Saturn. They also have a higher thrust-to-weight ratio. A consequence of these factors is a net gain of 18,144 kg (40,000 lbm) of payload compared to what could have been obtained using Saturn-type engines. The penalty for this performance is a considerably more complex engine that was difficult to develop and test.

The Orbiter is easily the largest man-made object to be recovered from space. Seventeen times heavier and with 50 times the surface area, it greatly exceeds the size of the Apollo command module, which previously held this distinction. Not surprisingly, the creation of a thermal protection system suitable for reuse after multiple reentries required a new material technology. Fundamentally, the choice lay between insulating the inner structure from a hot external skin or using external insulation that would keep the metallic skin cool enough to be part of the primary structure, as in conventional aircraft. The second approach was chosen as both lighter and less costly to produce. The cost savings were primarily associated with the straightforward use of conventional structural materials and manufacturing techniques commonly used in modern large airplanes. An equally important benefit was the large database from which the cost and weight of the major structural assemblies could be estimated. It also allowed additional time to analyze the thermal input because of the isolation of the structural and thermal protection system.

Except for a few regions in which the surface temperature will not exceed 644 K (700°F), the majority of the Orbiter's external surface is covered with lightweight tiles. There are more than 32,000 tiles, most of which are 15.2- or 20.3-cm (6 or 8 in) squares. The basic material of the tiles consists of a random matrix of fine quartz fibers. The compaction of the fibers and therefore the density is closely controlled during manufacture. Each tile is coated with a pigmented glaze that provides water proofing, abrasion resistance, and high heat radiation. The density of most of the tiles is comparable to that of balsa wood, although some tiles of a higher density are located in regions where expected physical loads or abuse may be greater. Because the tiles have a lower coefficient of expansion than the aluminum skin to which they are bonded, a felt pad is inserted between the tile and the skin to isolate strains due to the relative motion between skin and tile. The tiles should be suitable for 100-mission reuse if they never exceed a surface temperature of 1533 K (2300°F). If heating rates are higher, fewer reuses will be possible. One of the very good properties of the tile material is its capability of withstanding heating rates almost double its design value and still survive one flight. This was particularly reassuring during the first flight, for which heating rates could only be predicted based on a combination of theoretical analysis and wind-tunnel data. The worst case heating predictions fell well below the one-time capability of the material.

On 28 January 1986, the space shuttle, "Challenger," was destroyed by a failure of the seal on one of the solid rocket motors (SRM) that is intended to boost the vehicle into Earth orbit. Six astronauts and a schoolteacher acting as a "payload specialist" were on board "Challenger." All were killed. There is no doubt that this accident was the single most traumatic event in the effort to put people in space. The failure of the SRM occurred when the joint between the lowest and the next-to-lowest segment of the rocket stack failed and released the hot gases

inside the rocket (the location of this joint—called a “field joint” because it is made up in the field—is shown in Fig. 10). These hot gases rapidly melted the fixtures that attached the SRM to the fuel/oxidizer tank and caused the SRM to deflect and puncture the tank. The release of the hydrogen and oxygen quickly created a flammable mixture that ignited and destroyed the Orbiter.

A commission to investigate the accident chaired by former Secretary of State William P. Rogers was established. The commission found that the proximate cause of the accident was that the O-ring seal of the joint opened when subjected to the pressure of the gases inside the rocket. This was attributed to a design flaw in the O-ring seal that was exacerbated by the cold temperature on the day of the launch, which embrittled the O-ring material of the seal. The report of the commission also listed as possible contributing causes the problems encountered in the assembly of the rocket stack and the very high wind shear experienced at approximately the same time as the joint failed. The commission recommended that space shuttle flights be suspended until the seal of the joint was redesigned, manufactured, and tested so that failures of this kind would not be possible (14). It took more than 2 years to implement this recommendation, and Shuttle flight operations resumed in September 1988.

The “Challenger” was replaced by a new space shuttle, “Endeavor,” which flew for the first time in May 1992. Since September 1998, over one hundred space shuttle missions have been successfully executed. During these flights, many important in-orbit capabilities have been demonstrated, including the repair and retrieval of satellites by astronauts working outside the Space Shuttle cabin [extravehicular activities (EVA)], the execution of many onboard experiments

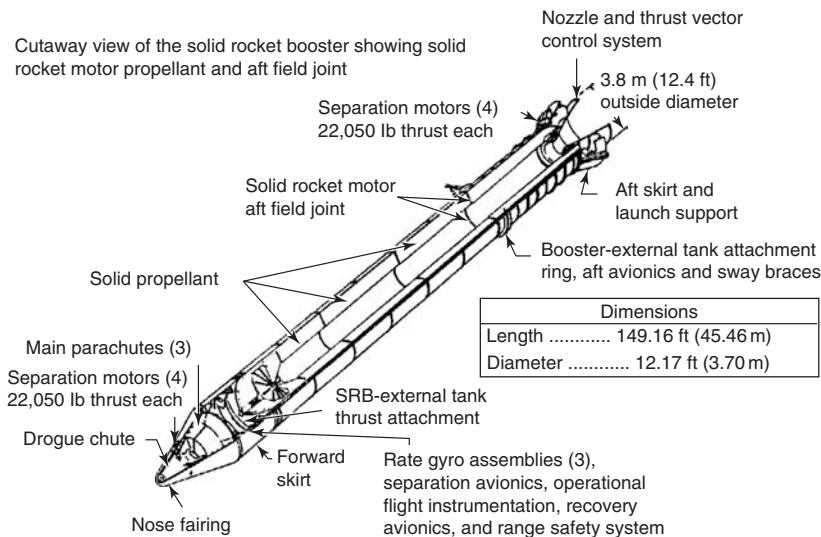


Figure 10. A cutaway drawing of the solid rocket motor of the Space Shuttle system. The joint that failed during the launch of “Challenger” is the “aft field joint” labeled on this drawing. The picture also shows the configuration of the rocket with the nozzle, the four segments of the solid fuel grain, the separation motors, and the location of the control system in the nose section of the rocket. This figure also can be seen at the following NASA website: <http://www.hq.nasa.gov/office/pao/History/rogersrep/v1p56.jpg>.

using the European Spacelab modules that were built for the Shuttle by the German space agency, docking with the Russian MIR space station, and launching important scientific satellites. The Space Shuttle fleet is currently providing the principal means for lifting many of the component parts of the International Space Station and all of the people who will assemble the station into Earth orbit. The Shuttles "Atlantis," "Discovery," and "Endeavor" have been uprated to deliver payloads to the higher inclination orbits of the International Space Station. The main engine is being modified and provisions such as docking hardware are being added. The external tank has also been redesigned primarily using a lithium aluminum material to increase payload capability. When the International Space Station is completed, the Space Shuttle will, for the first time, actually become what it has been called for so many years: the vehicle that will "shuttle" people and supplies back and forth to the International Space Station.

The Space Shuttle represents a bold major development in the technology of human spaceflight. It has made possible activities and achievements in space that would not have been done otherwise. It has been and will continue to be the major new stepping-stone to a new era in space-operations.

Conclusions

The enterprise of putting humans in space has been controversial from the very beginning. Shortly after the Mercury project was initiated in 1959, President Kennedy's science advisor, Dr. Jerome B. Wiesner, recommended giving serious consideration to canceling the program. In a report he submitted to President Kennedy shortly before his inauguration in January 1961, Dr. Wiesner argued that human spaceflight was very dangerous and could damage the new administration's prestige if some people were killed by accidents in space. Furthermore, he said that putting people in space is an expensive proposition and that the new administration's scientific and military objectives could best be achieved using robotic spacecraft. There is a kernel of truth in this argument, which is why it is still being heard now even after four decades of successful human spaceflight. What the people who make this argument either forget or ignore is the public impact of human spaceflight. The world craves heroes, and there is no doubt at all that our astronauts are modern heroes. John Glenn is a recent example. He was the first American to orbit the Earth in 1962 and flew again on the Space Shuttle, "Discovery," in 1999 when he was 77 years old. Successive administrations have been advised by very prestigious people to abandon the program to put people in space. Fortunately, this advice has been roundly ignored, and successive Presidents since John Kennedy have strongly supported the program to put people in space.

A good case can be made that 500 years from now, the only thing that will be remembered about the twentieth century is that on 20 July 1969, human beings first set foot on another body in our solar system. The lunar landing in 1969 is likely to be remembered just the way 1492 is the only year in the fifteenth century that the vast majority of people around the world remember. Christopher Columbus' landing on an island in the Bahamas has had the same impact as Neil Armstrong's landing on the Moon. Both of these events opened new horizons for human beings to aim at, and this is why they are both historically significant.

Appendix A: Pertinent Size and Performance Comparisons of U.S. Spacecraft

Item	Spacecraft	Mercury	Gemini	Apollo		Post-Apollo		Space Shuttle
				Command (service) modules(s)	Lunar module	Skylab	Apollo Soyuz	
Weight (Entry) kg (lbm)		1208 (2663)	2165 (4774)	5668 (12,497)	NA	6088 (13,425)	5843 (12,884)	90,249 (199,000) 71.46 (2525)
Volume (Habitable) m ³ (h ³)		1.02 (36)	1.56 (55)	5.94 (210)	4.53 (160)	Command module 5.94 (210) Skylab 344.98 (12,190)	Command module 5.94 (210) Docking module 19.81 (700)	
Duration (Maximum) days		1 1/2	13 3/4	12 1/2	3	84	9	7 to 30
Crew size		1	2	3	2	3/5	3	2 to 7
Cabin atmosphere mmHg (psi)		100% O ₂ at 258 (5)	100% O ₂ at 258 (5)	100% O ₂ at 258 (5)	100% O ₂ at 258 (5)	Command module 100% O ₂ at 258 (5)	Command module ^a 100% O ₂ at 222 (4.3) to 296 (5.6)	21% O ₂ /79% N ₂ at 760 (14.7)
						Orbital workshop 74% O ₂ / 26% N ₂ at 258 (5)	Docking module ^a 60% O ₂ / 40% N ₂ at 222 (4.3) to 520 (10)	

Suit usage	Cabin backup	Cabin backup	Cabin backup	Cabin backup	Cabin backup	Cabin backup
		Ejection	Extra vehicular activity	Crew transfer	Crew transfer	Crew transfer
		Extra vehicular activity	Crew transfer	Lunar surface excursion	Extra vehicular activity	Rescue
						Extra vehicular activity
						Rescue Ejection (OFT)
Propulsion main maneuvering and retro, Δv m/sec (ft/sec)	Solid retro	Solid retro	Service propulsion system	Descent propulsion system	Service propulsion system	Orbital maneuvering system
	98.8 (324)	99.1 (325)	195.1 (6401)	2,135 (7006)	533.1 (1749)	304.8 (1000)
				Ascent propulsion system		
				1850 (6070)		
Propulsion, reaction control system for auxiliary maneuvers and attitude control (total impulse) newton-sec (ft – sec)	30.967 (6.900)	Entry vehicle 90,478 (20,160)	Command module 256,714 (57,200)	782,483 (174,350)	Command module 329,531 (73,425)	9,236,304 (2,058,000)
		Orbital maneuvering system 1,077,524 (240,090)	Service module 1,653,828 (368,500)		Service module	
Lift/drag (Entry)	Ballistic	0.17 to 0.09 (Mach 24 to 6)	0.28 to 0.38 (Mach 36 to 6)	NA	(Same as Apollo)	1.90
Wetted area m ² (ft ²)	12.44 (133.9)	19.61 (211.1)	35.97 (387.2)	NA	(Same as Apollo)	1034.5 (11,136)
W/S kg/m ² (lbm/ft ³)	453.1 (93.0)	527.8 (108.11)	471.7 (96.6)	NA	(Same as Apollo)	361.3 (74.0)

^aO₂ Minimum partial pressure 165 mmHg (3.2 psi), N₂ variable during crew transfer.

Appendix B: Summary of American Manned Spaceflights

Spacecraft	Launch date	Crew	Flight time, hr:min:sec	Comments
<i>Mercury</i>				
MR-3	5 May 1961	Shepard	00:15:22	First suborbital flight
MR-4	21 July 1961	Crissom	00:15:37	Second suborbital flight
MA-6	20 Feb. 1962	Glenn	04:55:23	First Mercury orbital flight
MA-7	24 May 1962	Carpenter	04:56:05	Second Mercury orbital flight
MA-8	3 Oct. 1962	Schirra	09:13:11	Extended duration
MA-9	15 May 1963	Cooper	34:19:46	Extended duration
<i>Gemini</i>				
GT-3	23 Mar. 1965	Grissom/Young	04:52:31	First Gemini flight
GT-4	3 June 1965	McDivitt/White	97:40:01	20-min EVA, handheld maneuvering unit
GT-5	21 Aug. 1965	Cooper/Conrad	190:55:14	Extended duration
GT-7	4 Dec. 1965	Borman/Lovell	330:35:01	14-day mission
GT-6A	15 Dec. 1965	Schirra/Stafford	25:51:21	Rendezvous with GT-7
GT-8	16 Mar. 1966	Armstrong/Scott	10:41:26	Docked with Agena, Gemini reaction control system (RCS) malfunction: mission terminated
GT-9A	3 June 1966	Stafford/Cernan	72:20:50	Rendezvous with damaged Agena; 2-hr 7-min EVA
GT-10	19 July 1966	Young/Collins	70:46:39	Docked with Agena, 3 EVA periods
GT-11	12 Sept. 1966	Conrad/Cordon	71:17:08	Reached record altitude of 1373 km (341.5 nautical miles) using Agena propulsion
GT-12	11 Nov. 1966	Lovell/Aldrin	94:34:31	Docked with unusable Agena
<i>Apollo</i>				
Apollo 7	11 Oct. 1968	Schirra/Eisele/Cunningham	260:09:08	First Apollo light—Earth orbit
Apollo 8	21 Dec. 1968	Borman/Lovell/Anders	147:00:42	First manned lunar orbit, no LM
Apollo 9	3 Mar. 1969	McDivitt/Scott/Schweickart	241:00:53	First manned LM flight—Earth orbit

Apollo 10	18 May 1969	Stafford/Young/Cernan	192:03:23	LM and CSM rendezvous and docking activity in lunar orbit
Apollo 11	16 July 1969	Armstrong/Collins/Aldrin	125:18:35	First manned lunar landing, Sea of Tranquility, time on Moon (T.O.M.) 21:36:21
Apollo 12	14 Nov. 1969	Conrad/Gordon/Bean	244:36:25	Precise lunar landing near surveyor, Ocean of Storms, T.O.M. 31:31:12
Apollo 13	11 Apr. 1970	Lovell/Swiger/Haise	142:54:41	Lunar flyby during emergency return after service module damaged
Apollo 14	31 Jan. 1971	Shepard/Rees/Mitchell	216:01:58	Third landing—Fra Mauro, “golf cart” first wheels on Moon; T.O.M. 33:30:31
Apollo 15	26 July 1971	Scott/Worden/Irwin	295:11:53	Hadley-Apennine region, used Rover—an electric-powered car; T.O.M. 66:54:53
Apollo 16	16 Apr. 1972	Young/Mattingly/Duke	265:51:05	Descartes region, traveled 27 km in Rover, T.O.M. 71:02:13
Apollo 17	6 Dec. 1972	Cernan/Evans/Schmitt	301:51:59	Taurus-Littrow region, traveled 35 km in Rover, T.O.M. 14:59:38
<i>Skylab</i>				
Skylab 2	25 May 1973	Conrad/Kerwin/Weicz	672:49:49	Repaired damage to Skylab; initial space experiment activity
Skylab 3	28 July 1973	Bean/Garriott/Lousma	1427:09:04	Increased duration; continued experiment activity
Skylab 4	16 Nov. 1973	Carr/Gibson/Pogue	2017:15:31	Increased duration; completed experiment activity
<i>Apollo-Soyuz Test Project</i>				
ASTP-Apollo	15 July 1975	Stafford/Brand/Slayton	217:28:24	Rendezvous and docking with Soyuz
<i>Space Shuttle</i>				
STS	Launch Date	Crew	Shuttle	Mission
1	12–14 Apr. 1981	Young/Crippen	Columbia	Successful first flight
2	12–14 Nov. 1981	Engle/Truhy	Columbia	Test of remote manipulator, payload bay experiments
3	22–30 Mar. 1982	Lousma/Fullerton	Columbia	Space viewing experiments, only landing at White Sands

Appendix B: (Continued)

Spacecraft	Launch date	Crew	Flight time, hr:min:sec	Comments
4	27 June-4 July 1982	Mattingly/Hartsfield	Columbia	Last test flight. Carried DOD payloads. First runway landing at Edwards AFB
5	11-16 Nov. 1982	Brand/Overmyer/Allen/Lenoir	Columbia	Two commercial satellites SBS-3 and Anik C-3
6	4-9 Apr. 1983	Weitz/Bobko/Peterson/Musgrave	Challenger	Tracking and Data Relay Satellite (TDRS); first spacewalk
7	18-24 June 1983	Crippen/Hauck/Ride/Fabian/Thagard	Challenger	Two commercial satellites Anik- and Palapa B-1. Deployed and retrieved SPAS. First flight of American woman.
8	30 Aug.-5 Sep. 1983	Truly/Brandenstein/Bluford/Gardner/Thorton	Challenger	Communication satellites, India's INSAT 1B. First flight of African-American
9	28 Nov.-8 Dec. 1983	Young	Challenger	Spacelab module flight. First non-U.S. crew member
41B	3-11 Feb. 1984	Brand/Gibson/McCandless/Stewart/NcNair	Challenger	Communication satellites Westar VI and Palapa B-2. Both satellite propulsion systems failed. First flight of manned maneuvering unit. First landing at Kennedy Space Center
41C	6-13 Apr. 1984	Crippen/Scobee/Hart/van Hooten/Nelson	Challenger	Long duration exposure facility. Rendezvous and repair of Solar Maximum satellite
41D	30 Aug.-5 Sep. 1984	Hartsfield/Coats/Mullane/Hawley/Resnik/Walker	Discovery	Communication satellites Syncom IV-2, SBS-4 and Telstar 3C. Experiment with folding solar array. First commercial payload specialist crew member
41G	5-13 Oct. 1984	Crippen Mc Bride	Challenger	Earth observation experiments. First Canadian crew member
51A	8-16 Nov. 1984	Hauck/Walker/Fisher/Gardner	Discovery	Communication satellites Anik D2 and Syncom IV-1. Recovered failed Palapa B-2 and Westar VI
51C	24-27 Jan. 1985	Mattingly/Shiver/Onizuka/Buchli/Payton	Discovery	DOD payload
51D	12-19 Apr. 1985	Bobko/Williams/Hoffman/Griggs/Seddon/Walker/Garn	Discovery	Communication satellite Anik C-1 and Syncom IV-3. Syncom IV-3 propulsion system failed to ignite. First Congressional crew member, Jake Garn

51B	29 Apr.–6 May 1985	Overmeyer/Gregory/Lind/ Thagard/Thorton/Wang/ van der Berg	Challenger	Spacelab mission
51G	17–24 June 1985	Brandenstein/Creighton/ Fabian/Nagel/Lucid/ Bandry/Al Sa'ud	Discovery	Commercial satellites Morelos-1, Arabsat 1-B and Telstar 3D
51F	29 July–6 Aug. 1985	Fullerton/Bridges/ Musgrave/England/ Henize/Action/Bartoe	Challenger	Spacelab 2 with open pallet
51I	27 Aug.–3 Sep. 1985	Engle/Covey/van Hoften/ Fisher/Lounge	Discovery	Communications satellites ASC-1, AUSSAT-1 and Syncom IV-4. Repaired and reboosted Syncom IV-3 which failed on flight 51D
51J	3–7 Oct. 1985	Bobko/Grabe/Hilmers/ Stewart/Pailes	Atlantis	DOD payload
61A	30 Oct.–6 Nov. 1985	Hartsfield/Nagel/Buchi/ Bluford/Dunbar/Furrer/ Ockel/Messerschmid	Challenger	Spacelab D-1
61B	26 Nov.–3 Dec. 1985	Shaw/O'Connor/Spring/ Cleave/Ross/Walker/Vela	Atlantis	Communication satellites SATCOM Ku 2 Morelos 2 and AUSSAT-2. Construction experiment. First Mexican crew member
61C	12–18 Jan. 1986	Gibson/Bolden/Nelson/ Hawley/Chang-Diaz/ Cenker/Nelson	Columbia	RCA satellite
51L	28 Jan. 1986	Scobee/Smith/Onizuka/ Resnik/Mc Nair/Jaris/ McAuliffe	Challenger	Failed
26	29 Sep.–3 Oct. 1988	Hauck/Covey/Lounge/ Nelson/Hilmers	Discovery	TDRS satellite
27	2–6 Dec. 1988	Gibson/Gardner/Mullone/ Ross/Shepherd	Atlantis	DOD payload
29	13–18 Mar. 1989	Coats/Blaha/Buchli/ Springer/Bagian	Discovery	TDRS satellite
30	4–8 May 1989	Walker/Grabe/Thagard/ Cleave/Lee	Atlantis	Magellan Venus Radar Mapper

Appendix B: (Continued)

Spacecraft	Launch date	Crew	Flight time, hr:min:sec	Comments
28	8–13 Aug. 1989	Shaw/Richards/Leestma/ Adamson/Brown	Columbia	DOD payload
34	18–23 Oct. 1989	Williams/Meculley/Lucid/ Baker/Chang-Diaz	Atlantis	Galileo (Jupiter)
33	22–27 Nov. 1989	Gregory/Blaha/Musgrave/ Thornton/Carter	Discovery	DOD payload
32	9–20 Jan. 1990	Brandenstein/Wetherbee/ Dunbar/Low/Ivins	Columbia	Communication satellite SYNCOM LDEF recovery
36	8 Feb.–4 Mar. 1990	Creighton/Casper/Hilmers/ Mullane/Thuot	Atlantis	DOD payload
31	24–29 Apr. 1990	Shriver/Bolden/Hawley/ McCandless/Sullivan	Discovery	Hubble Space Telescope
41	6–10 Oct. 1990	Richards/Cabana/Melnick/ Shepherd/Akers	Discovery	ESA Ulysses (Sun)
38	15–20 Nov. 1990	Covey/Culbertson/Springer/ Meade/Gemar	Atlantis	DOD payload
35	2–10 Dec. 1990	Brand/Gardener/Hoffman/ Lounge/Parker/Parise/ Durrance	Discovery	Spacelab astrophysics
37	5–11 Apr. 1991	Nagel/Cameron/Apt/ Godwin/Ross	Atlantis	Gamma Ray Observatory
39	28 Apr.–6 May 1991	Coats/Hammond/Bluford/ Harbaugh/Hieb/ Mcmonagle/Veach	Discovery	DOD payload
40	5–14 June 1991	O'Connor/Gutierrez/Bagian/ Jernigan/Seddon/Gaffney/ Fulford	Columbia	Spacelab life sciences
43	2–11 Aug. 1991	Blaha/Baker/Adamson/Low/ Lucid	Atlantis	TDRS-5

48	12-18 Sep. 1991	Creighton/Reightler/Brown/ Gemar/Buchli	Discovery	Upper Atmosphere Research satellite
44	24 Nov.-1 Dec. 1991	Gregory/Henricks/Runco/ Voss/Musgrave/Hennen	Atlantis	DOD payload
42	22-30 Jan. 1992	Grabe/Oswald/Readdy/ Thagard/Hilmers/ Bondar/Merbold	Discovery	Spacelab International Microgravity Laboratory
45	24 Mar.-2 Apr. 1992	Bolden/Duffy/Sullivan/ Leestna/Foale/Frimout/ Lihtenberg	Atlantis	Spacelab Atmospheric Laboratory for Applications and Science
49	7-16 May 1992	Brandenstein/Chilton/ Melnick/Akers/Hieb/ Thornton/Thuot	Endeavour	Spacewalks and repair and redeployment of INTELSAT VI
50	25 June-9 July 1992	Richards/Bowersox/Dunbar/ Meade/Baker/DeLucas/ Trinh	Columbia	Microgravity Laboratory
46	31 July-8 Aug. 1992	Shriver/Allen/Hoffman/ Chang-Diaz/Ivins/ Nicollier/Malerba	Atlantis	ESA European Retrievable Carrier, EURECA and NASA-Italian Tethered Satellite System
47	12-20 Sep. 1992	Gibson/Brown/Lee/Davis/ Jemison/Mohri	Endeavour	Spacelab J (Japanese Spacelab). First Japanese crew member
52	22 Oct.-1 Nov. 1992	Wetherbee/Baker/Veatch/ Jernigan/Shepard/ MacLean	Columbia	Laser Geodynamic Satellite II (LAGEOS)
53	2-9 Dec. 1992	Walker/Cabana/Bluford/ Voss/Clifford	Discovery	DOD payload
54	13-19 Jan 1993	Casper/McMonagle/Runco/ Harbaugh/Helms	Endeavour	TDRS-6
56	8-17 Apr 1993	Cameron/Oswald/Cockrell/ Foale/Ochos	Discovery	ATLAS_Mission to Planet Earth
55	26 Apr-6 May 1993	Nage/Henricks/Ross/ Precourt/Harris/Walter/ Schlegel	Columbia	German Spacelab D-2

Appendix B: (Continued)

Spacecraft	Launch date	Crew	Flight time, hr:min:sec	Comments
57	21 Jun.-1 Jul. 1993	Grabe/Duffy/Low/Sherlock/ Voss/Wisoff	Endeavour	First Spacehab
51	12-22 Sep. 1993	Culbertson/Readdy/ Newman/Bursh/Wakz	Discovery	Advanced Communication Technology Satellite (ACTS) and Retrievable Far and Extreme Ultraviolet Spectrograph (ORFEUS)
58	18 Oct.-1 Nov. 1993	Blaha/Searfoss/Secdon/ McArthur/Wolf/Lucid/ Fettman	Columbia	Spacelab Life Science Mission
61	2-13 Dec. 1993	Covey/Bowersox/Musgrave/ Hoffman/Thornton/Akers/ Nicollier	Endeavour	Servicing of Hubble Space Telescope
60	3-11 Feb. 1994	Bolden/Reightler/Chang- Diaz/Davis/Sega/Krikalev	Discovery	SPACEHAB-2. First Russian crew member
62	4-18 Mar. 1994	Casper/Allen/Gemar/Ivins/ Thuot	Columbia	Microgravity Payload-2 (USMP-2) and Office of Aeronautic and Space Technology (OAST)
59	9-20 Apr. 1994	Gutierrez/Chilton/Godwin/ Apt/Clifford/Jones	Endeavour	Space Radiation Laboratory-2
65	8-23 July 1994	Cabana/Halsell/Hieb/ Thomas/Walz/Chiao/ Naito-Mukai	Columbia	Microgravity Laboratory-2
64	9-20 Sept. 1994	Richards/Hammond/Helms/ Meade/Lee/Linenger	Discovery	Lidar In-Space Technology Experiment (LITE)
68	30 Sept.-11 Oct. 1994	Baker/Wilcutt/Jones/Bursh/ Wisoff/Smith	Endeavour	Space Radiation Laboratory-2
66	3-14 Nov. 1994	McMonagle/Brown/Ochoa/ Tanner/Paraznynski/ Clervoy	Atlantis	Atmospheric Laboratory for Applications and Science (ASTRO 3 and Cryogenic Infrared Spectrometer and Telescope for the Atmosphere (CRISTA)-SPAS
63	3-11 Feb. 1995	Wetherbee/Collins/Harris/ Foale/Voss/Titov	Discovery	SPACEHAB-3, SPARTAN-204 satellite. MIR rehearsal

67	2–18 Mar. 1995	Oswald/Gregory/Grunsfeld/ Lawrence/Jernigan/ Parise/Durrance	Endeavour	Atmospheric Laboratory for Applications and Science ASTRO 2
71	27 Jun.–7 Jul. 1995	Gibson/Precourt/Baker/ Harbaugh/Dunbar	Atlantis	First MIR mission
70	13–22 Jul. 1995	Henricks/Kregel/Currie/ Thomas/Weber	Discovery	TDRS-7
69	7–18 Sept. 1995	Walker/Cockrell/Voss/ Newman/Gernhardt	Endeavour	Wake Shield Facility and Spartan spacecraft
73	20 Oct.–5 Nov. 1995	Bowersox/Rominger/ Thornton/Coleman/ Lopez-Alegria/Leslie/ Sacco	Columbia	Microgravity Laboratory-2
74	12–20 Nov. 1995	Cameron/Halsell/Hadfield/ Ross/McArthur	Atlantis	Second MIR mission
72	11–20 Jan. 1996	Duffy/Jett/Chiao/Barry/ Scott/Wakata	Endeavour	Capture and return Microgravity Research Spacecraft
75	22 Feb.–9 Mar. 1996	Allen/Horowitz/Hoffman/ Cheli/Nicollier/Chang- Diaz/Guidoni	Columbia	Reflight of Tethered Satellite
76	22–31 Mar. 1996	Chilton/Searfoss/Godwin/ Sega/Clifford/Lucid	Atlantis	Third MIR mission
77	19–29 May 1996	Casper/Brown/Bursch/ Runco/Garneau/Thomas	Endeavour	SPACEHAB
78	20 Jun.–7 Jul. 1996	Henricks/Kregel/Helms/ Linnehan/Brady/Favier/ Tirsk	Columbia	Spacelab Life and Microgravity flight
79	16–26 Sep. 1996	Readdy/Wileutt/Akers/Apt/ Walz/Blaha/Lucid	Atlantis	Fourth Mir mission
80	19 Nov.–7 Dec. 1996	Cockrell/Rominger/ Jernigan/Jones/Musgrave	Columbia	Third flight of Wake Shield Facility and third flight of German ORFEUS SPAS II
81	12–22 Jan. 1997	Baker/Jett/Wisoff/ Grunsfeld/Blaha	Atlantis	Fifth MIR mission

Appendix B: (Continued)

Spacecraft	Launch date	Crew	Flight time, hr:min:sec	Comments
82	11–21 Feb. 1997	Bowersox/Horowitz/Tanner/ Hawley/Harbaugh/Lee/ Smith	Discovery	Hubble Space Telescope serving
83	4–8 Apr. 1997	Halsell/Still/Voss/ Gernhardt/Thomas/ Crouch/Linteris	Columbia	Microgravity Science Laboratory-1. Mission terminated after 4 days due to fuel cell problem
84	15–24 May 1997	Precourt/Collins/Clervoy/ Noriega/Lu/Kondakova/ Foale/Linegar	Atlantis	Sixth Mir mission
94	1–17 Jul. 1997	Halsell/Still/Voss/ Gernhardt/Thomas/ Crouch/Linteris	Columbia	Microgravity Science Laboratory MSL-1 refight of STS 83
85	7–19 Aug. 1997	Brown/Rominger/Davis/ Curbeam/Robinson/ Tryggvason	Discovery	CRISTA-SPAS-2
86	25 Sep.–6 Oct. 1997	Wetherbee/Bloomfield/ Parazynski/Titov/ Chretien/Lawrence/Wolf/ Foale	Atlantis	Seventh MIR mission
87	19 Nov.–5 Dec. 1997	Kregel/Lindsey/Chawla/ Scott/Doi/Kadenyuk	Columbia	Experiments of weightless environment
89	22–31 Jan. 1998	Wilcutt/Edwards/Reilly/ Anderson/Dunbar/ Shakirovich/Thomas/Wolf	Endeavour	Eighth MIR mission
90	17 Apr.–3 May 1998	Searfoss/Altman/Hire/ Linnehan/Williams/ Buckey/Pawelczyk	Columbia	Spacelab 16th and final flight
91	2–12 June 1998	Precourt/Gorie/Lawrence/ Chang-Diaz/Kavandi/ Thomas/Ryumin	Discovery	Ninth and last MIR mission

95	29 Oct.-7 Nov. 1998	Brown/Lindsey/Parazynski/ Robinson/Duque/Mukai/ Glenn	Discovery	Space lab, Glenn return to flight
88	4-15 Dec. 1998	Cabana/Sturckow/Newman/ Currie/Ross/Krikalev	Endeavour	First flight to ISS Space Station with Unity module
96	27 May-6 June 1999	Rominger/Husband/Barry/ Tokarev/Ochoa/Payette/ Jernigan	Discovery	ISS supply
93	23-27 July 1999	Collins/Ashby/Tognini/ Hawley/Coleman	Columbia	Chandra X-Ray Observatory
103	19-26 Dec. 1999	Brown/Kelly/Smith/Foale/ Grunsfeld/Nicollier/ Clervoy	Discovery	Hubble space Telescope servicing #3
99	11-22 Feb. 2000	Kregel/Gorie/Thiele/ Kavandi/Voss/Mohri	Endeavour	Radar Topograph
101	19-28 May 2000	Halsell/Horowitz/Helms/ Usachev/Voss/Weber/ Williams	Atlantis	Third ISS
106	8-19 Sep. 2000	Wilcutt/Altman/Burbank/ Lu/Malenchenko/ Mastiacchio/Morukov	Atlantis	Fourth ISS
92	11-24 Oct. 2000	Lopez-Alegria/Duffy/ Melroy/McArthur/Chiaco/ Wisoff/Wakata	Discovery	Fifth ISS
97	30 Nov.-11 Dec. 2000	Bloomfield/Jett/Garneau/ Noriega/Tanner	Endeavour	Sixth ISS
98	7-20 Feb. 2001	Cockrell/Polansky/ Curbeam/Ivins/Vones	Atlantis	Seventh ISS
102	8-21 Mar. 2001	Wetherbee/Kelly/Thomas/ Richards/Krikalev/Voss/ Shephard/Helms/ Gidzenko/Usachev	Discovery	Eighth ISS Crew return

Appendix B: (Continued)

Spacecraft	Launch date	Crew	Flight time, hr:min:sec	Comments
100	19 Apr–1 May 2001	Rominger/Ashby/Hadfield/ Phillips/Parazynski/ Guidoni/Lonchakov	Endeavour	Ninth ISS
104	12–24 Jul 2001	Lindsey/Habaugh/ Gernhardt/Kavandi/ Reilly	Atlantis	Tenth ISS
105	10–22 Aug. 2001	Horowitz/Sturckow/ Forrester/Barry/Voss/ Usachev/Helms/ Dezhurow/Culbertson/ Tyurin	Discovery	Eleventh ISS Crew return
108	5–17 Dec. 2001	Gorie/Kelly/Goodwin/Tani/ Onufrienko/Walz/Bursch/ Culbertson/Turin/ Dezhurow	Endeavour	Twelfth ISS
109	1–12 Mar. 2002	Altman/Carey/Grunsfeld/ Currie/Newman/ Linnehan/Massimino	Columbia	Hubble Service
111	5–19 Jun. 2002	Cockrell/Cockhart/Chang- Diaz/Perrin/Korzum/ Whitson/Treschev/ Onufriyenko/Walz/ Bursch	Endeavour	Thirteenth ISS
112	7–18 Oct. 2002	Ashby/Melroy/Wolf/Sellers/ Magnus/Yurchikhin	Atlantis	Fourteenth ISS

BIBLIOGRAPHY

1. Swenson, L.S., J.M. Grimwood, and C. Alexander. *This New Ocean: A History of Project Mercury*.
2. Hacker, B.C., and J.M. Grimwood. *On the Shoulders of Titans: A History of Project Gemini*.
3. McDougall, W.A. *The Heavens and the Earth: On the Shoulders of Titans: A History of Project Gemini*, NASA SP-4203, U.S. Government Printing Office, Washington, DC, 1977.
4. Sidey, H. *Time Magazine*, Nov. 14, 1983.
5. Ertel, I.D., and M.L. Morse. *The Apollo Spacecraft: A Chronology*. NASA SP-4009, US Government Printing Office, Washington, DC, 1969.
6. Bilstein, R.E. *Stages to Saturn: A Technological History of the Apollo/Saturn Launch Vehicles*. NASA SP-4206, U.S. Government Printing Office, Washington, DC, 1980.
7. Hansen, J.R. *Spaceflight Revolution: NASA Langley Research Center from Sputnik to Apollo: A History of Manned Lunar Spacecraft*. NASA SP-4205, U.S. Government Printing Office, Washington, DC, 1995.
8. Brooks, C.G., J.M. Grimwood, and L.S. Swenson, Jr. *Chariots for Apollo: A History of Manned Lunar Spacecraft*. NASA SP 4205, U.S. Government Printing Office, Washington, DC, 1979.
9. Lovell, J., and J. Kluger. *Lost Moon: The Perilous Voyage of Apollo 13*. Houghton Mifflin, Boston, New York, 1994.
10. Compton, W.D., and C.D. Benson. *Living and Working in Space: A History of Skylab*. NASA SP-4208, U.S. Government Printing Office, Washington, DC, 1983.
11. Foehlich, W. *Apollo-Soyuz*. NASA EP-109, U.S. Government Printing Office, Washington, DC, 1976.
12. Heppenheimer, T.A. *The Space Shuttle Decision: NASA's Search for a Reusable Space Vehicle*. NASA SP-4221, U.S. Government Printing Office, Washington, DC, 1999.
13. Chaisson, E.J. *The Hubble Wars*. Harper Collins, New York, 1994.
14. Report of the Presidential Commission on the Space Shuttle "Challenger" Accident. Washington, DC, June 6, 1986, Vol. I.

MAXIME A. FAGET
MILTON A. SILVEIRA
Formerly with the Engineering and
Development Directorate
NASA-Johnson Space Center
Houston, Texas

URANUS AND NEPTUNE

Uranus and Neptune, were the first planets discovered by telescopic observation. They are Jovian planets of similar size and much more massive than Earth (by factors of 14.5 and 17) but are still far less massive than Jupiter (by factors of 22 and 18). Both have deep atmospheres dominated by hydrogen, but unlike Jupiter and Saturn, most of their mass consists of rocky and icy components, which are

in fluid form due to high interior temperatures. Both have ring systems, a diverse system of satellites, unusual tilted and offset magnetic fields, and similar atmospheric circulations. Their blue colors distinguish them from the pale tans of Jupiter and Saturn. But they also differ significantly from each other; they have very different obliquities (spin axis inclinations relative to their orbital planes), vastly different internal heat fluxes, and remarkably different weather patterns. Their orbital and physical parameters are summarized in Table 1.

Discovery

Uranus, at visual magnitude 5.5, is about as bright as Jupiter’s brightest moon and is close to the limit of visual detectability (magnitude 6). Neptune (magnitude 7.85) is far too dim (by a factor of 30) to be seen by a visual observer. Thus, it is not surprising that the discovery of both planets awaited the development of advanced telescopes. Although observed as early as 1690 by Flamsteed (4) and assumed to be a star, it was not until 13 March 1781 that Uranus was discovered by William Herschel, the best telescope maker of his time. Its extended image

Table 1. **Orbital and Physical Parameters of Uranus and Neptune**

	Uranus	Neptune
Orbital Parameters:		
Mean orbital radius, AU ^{a,d}	19.19126	30.06896
Orbital eccentricity ^{a,e}	0.04717	0.00859
Orbital sidereal Period (Y) from ssd.jpl.nasa.gov (May 22, 2001 update)	84.017	164.791
Orbital inclination to ecliptic ^{a,f}	0.77°	1.77°
Planetary physical parameters:		
Mass ^g (Earth masses, Earth = 5.974 × 10 ²⁷ g)	14.535	17.141
Equatorial radius at 1 bar level, km ^b	25,559	24,766
Polar radius at 1 bar level, km ^b	24,973	24,342
Ellipticity (R _{equator} /R _{pole} – 1)	0.0229	0.0171
Obliquity (tilt of rotational pole to orbital pole) ^c	97.86°	29.56°
Sidereal rotation period of interior ^b	17.240 h	16.11 h
Mean density ^g , g/cm ³	1.318	1.638
Equatorial surface gravity, m/s ² , 1 bar level ^g (Earth = 9.78)	8.69	11.00

^aRef. 1., p. 316, Table 5.8.1.
^bRef. 2.
^cHere the rotational pole and orbital pole vectors are both defined by the right hand rule: the pole direction is such that when the pole vector points at the observer, the observer will see counter-clockwise motion. In the case of Uranus, the rotational pole points opposite to the planet’s North Pole, as defined by the International Astronomical Union (IAU). Atmospheric dynamicists often find it more convenient to treat the rotational pole as the North Pole.
^dOne AU = 149.60 × 10⁶ km.
^eEccentricity = distance between ellipse foci divided by 2 × semimajor axis.
^fEcliptic plane = plane of earth’s orbit.
^gRef. 3.

implied to Herschel that it was not a star, though at first he believed that it was a comet. Herschel was an amateur at the time, but later became the first president of the Royal Astronomical Society. This was the first planet discovery that can be attributed to a specific individual at a particular time. All of the brighter planets were known to the ancients. Neptune's discovery was even more unusual. It was the first planet discovered with the aid of mathematical predictions. Prediscovery observations and 40 years of post discovery observations of Uranus showed that its motion was being perturbed from its expected elliptical orbital path. John Couch Adams of England and Urbain Jean Joseph Le Verrier of France were independently able to use these perturbations to infer the location of a previously unknown planet (Neptune). Le Verrier was the more successful in publishing his calculations and convincing an observer to look in his predicted direction (5). Following receipt of Le Verrier's coordinates, John Galle and his assistant, Heinrich D'Arrest of Germany, first located Neptune on 23 September 1846 and confirmed its position and disc-shaped image the following night, more than a year after the first prediction had been made, but only five days after Le Verrier sent his request to Galle. Various prediscovery observations of Neptune were subsequently identified, the earliest by Galileo in 1612 (5)!

Formation

According to current theories (6,7), the primordial nebula from which the solar system formed is comprised of 98% hydrogen and helium. In the part of the nebula where the outer planets formed, the remaining 2% was dominated by water, ammonia, methane, and rock, all of which were probably in condensed form in the vicinity of Uranus and Neptune. Most of the condensates were "ices," a term applied to methane and ammonia, as well as to water, even when they are not actually in solid form. About a quarter of the solids were rock. This very small fraction of condensed solids played the key role of accreting into asteroid-sized planetesimals (of the order of a kilometer in size), which subsequently accreted into planetary cores. When these cores grew to a critical mass, they were then able to attract significant amounts of nebular gas (mostly hydrogen and helium). The process probably proceeded more slowly for Uranus and Neptune than for Jupiter and Saturn due to lower nebular densities and slower orbital velocities, and thus they were able to capture less nebular gas before the nebula was dissipated by the intense solar wind that developed during the Sun's early evolution. Uranus and Neptune were thus formed by materials of a higher average density than the materials that formed Jupiter and Saturn. The rocky cores of Uranus and Neptune are each about one-sixth of the planet's radius and are surrounded by icy material out to 75–80% of the radius; the outer 20–25% is primarily a hydrogen and helium gaseous envelope. During accretion of planetesimals, a planet will gain angular momentum as well as mass. A planet in an eccentric orbit will gain the greatest angular momentum at the edges of the accretion zone and will tend to accumulate prograde angular momentum (8). During the formation of both Uranus and Neptune, the last accreted objects might have been relatively large and had sufficient angular momentum to shift the spin axis significantly away from its "natural" spin direction, which is

perpendicular to the ecliptic plane. The last large object that hit Uranus may have been the size of Earth and might have played a role in generating debris from which the satellites may have formed (8). The inclined angular momentum of the Neptune system is also a probable result of impacts; the largest impactor was between 0.1 and 0.5 Earth masses (9). When satellites form from debris created by a late large impactor, most of the debris material is from the impactor, which thus determines the satellite composition.

The regular satellites are those that have nearly circular orbits close to the equatorial plane of the planet. These must have formed after the final large impact that shifted the angular momentum of the planet, and they might have formed from the debris generated by that impact. Any existing material in orbit near the planet at the time of the impact would be perturbed into inclined orbits that would result in collisions and breakup, which would then also contribute to accretion into satellite systems. Irregular satellites that have highly inclined, retrograde orbits are indicative of captured objects. Neptune's largest satellite, Triton, is plausibly a captured object. It could have been captured during a close approach to Neptune if it had impacted a regular satellite of about 1% of its mass, which would then slow Triton sufficiently to keep it in orbit around Neptune (8).

Interior Structure

The planets' masses, shapes, rotational rates, and gravitational moments place constraints on their interior structures. Models of both planets can be constructed using three shells (10). Some models use an inner core of rocky material, an intermediate shell of icy material, and an outer layer of gas. The ratio of ice to rock in these models is about 15, and they require an atmosphere enhanced in volatiles by about a factor of 20 compared with solar fractions. Neptune's rock-ice boundary in these models is about 15–20% of the planet's radius. The ice-atmosphere boundary is about 80% of the planet's radius. Successful models have also been created by using a gradual transition between the ice shell and the hydrogen-rich outer shell. Considerable uncertainty remains concerning the composition of the deep Neptune atmosphere and of the size of the rocky core (it might be about one Earth mass or considerably smaller). It is thought that the icy shells on both Uranus and Neptune are largely chemically homogeneous and consist of a mixture of ice, rock, hydrogen, and helium (10). Compared to Neptune, Uranus is somewhat less dense and somewhat more centrally condensed. As shown in Fig. 1, this structure has three main differences from that of Jupiter. Jupiter has a much larger gaseous molecular envelope, a large region of metallic hydrogen, and a smaller volume of ices.

Neptune loses internal heat at a rate 30 times larger (in power/unit mass, or luminosity) than what could be provided by radioactive decay of trace elements (the Earth's internal heat source). Neptune's heat source, like Jupiter's, is thought to be primordial heat left from the process of formation. The capture of solid and gaseous materials generated heat that raised the interior temperatures to very high levels (thousands K). That interior heat is still being released into space by Neptune, which emits 2.6 times as much heat as it absorbs from the Sun (12). This means that internal heat loss contributes 1.6 times as much as the

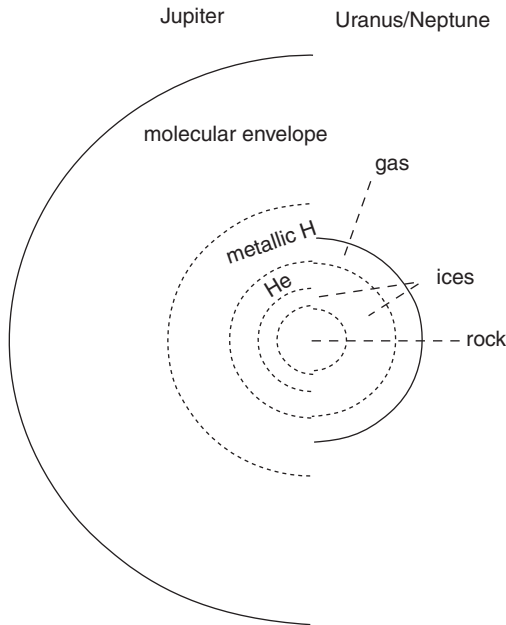


Figure 1. Approximate interior structures of Uranus and Neptune compared to Jupiter (11). Rocky and icy material may be mixed on Uranus and Neptune. The gas layer on Uranus is probably thicker than that on Neptune because Neptune is slightly denser. The radius at which metallic hydrogen is formed on Jupiter has been revised recently. Hubbard (7) now places it at about $0.8 R_J$.

reradiated heat absorbed from the Sun. It is somewhat of a mystery that no comparable heat emission is measured for Uranus, which is so similar to Neptune in most other respects. The lack of an internal heat source explains why Uranus' effective temperature (59 K) is the same as Neptune's even though it is much closer to the Sun (where sunlight is about 2.5 times as intense). Estimates of the heats of formation for Uranus and Neptune and the estimated cooling since formation suggest that the remaining primordial heat is more than sufficient to explain the present luminosity of Neptune. One explanation for the low luminosity of Uranus is that there are compositional gradients in the interior of Uranus that lead to suppressed convection (10), which reduces heat transfer efficiency and thus the temperature of its outer atmospheric layers. Such gradients might have arisen in connection with the giant impact that is presumably responsible for Uranus' spin axis inclination. An alternate theory (13) suggests that different external solar forcing produced by the high obliquity of Uranus has made the heat transfer process on Uranus much more efficient than on Neptune and resulted in a more rapid loss of internal heat.

The generation of an offset dipole magnetic field, that has a significant quadrupole component requires the existence of a conducting fluid layer in convection. This layer might be bounded at the deepest level by the point below which the interior is stably stratified, and thus not convecting, and the point above which the fluid interior is not electrically conducting. If the entire interior

fluid of Neptune exhibited the same angular rotation as that observed by clouds at the top of the atmosphere, then the J_4 gravitational expansion coefficient would be positive, whereas the observed value is negative. This implies that the differential rotation of the upper atmosphere is a superficial effect that does not involve a significant fraction of the planet's mass. However, this analysis does not place a well-defined boundary on how deep that flow could extend.

Atmospheres

Neptune and Uranus have similar compositions, similar tropospheric temperature structures, and similar styles of zonal circulation, but different cloud patterns and great differences in weather activity. The basic atmospheric parameters for Uranus and Neptune are summarized in Table 2.

Composition. Hydrogen, helium, and methane are the most prominent components of the atmospheres of Uranus and Neptune. Ammonia and probably H_2S and water are present in layers below the region accessible by optical spectroscopy. Hydrocarbons of various types are observed in the upper atmosphere, a result of UV-induced chemistry involving the breakdown of CH_4 .

The presence of hydrogen in any outer planet atmosphere was established first for Uranus, using the pressure induced 3-0 S(1) line of molecular hydrogen at 825.8nm, which was first measured by Kuiper (15), but first identified by Herzberg (16). Molecular hydrogen occurs in two forms, the ortho form in which the nuclear spin vectors of the two atoms are parallel, and the para form in which the nuclear spin vectors are antiparallel. At high temperatures, the two forms reach an equilibrium concentration of three ortho molecules to one para molecule. That is called normal hydrogen. Hydrogen convected to lower temperatures will retain the normal mixing ratio for a long time because of the very weak interaction between the two nuclei in the absence of a catalyst. Normal hydrogen was expected on Jovian planets because of rapid mixing that should overwhelm the slow conversion process. With an effective catalyst and sufficient time, the two forms will reach an equilibrium distribution that depends on temperature, and with sufficiently rapid conversion, large effects on specific heat, atmospheric buoyancy, and temperature lapse rate can occur. When conversion is slow, the

Table 2. **Atmospheric Composition^a**

	Uranus	Neptune
Atmospheric composition (percent of total volume or molecular number)		
H ₂ (Sun = 84)	83	79
He (Sun = 16)	15	18
H ₂ O (solar O = 0.15)	?	?
CH ₄ (solar C = 0.07)	2 (30 × solar)	3 (40 × solar)
NH ₃ (solar N = 0.02)	?	?
H ₂ S (solar S = 0.003)	?	?

^aRef. 14.

two forms behave like independent gases, and the specific heat at constant pressure (C_p) is lower. This leads to a steeper adiabatic lapse rate, given by $dT/dz = g/C_p$, where g is the local gravitational acceleration. Observational constraints are conflicting (17). The hydrogen quadrupole spectrum and the collision-induced dipole spectrum measured in the infrared are both approximately consistent with thermal equilibrium for the two forms of hydrogen at the temperature at which the spectral lines are formed. But the measured temperature lapse rate is more consistent with that expected for normal hydrogen. This conflict might be resolved with the concept of stratified convection layers (18) in which a given gas parcel resides in a thin layer long enough to reach ortho-para equilibrium, even though the convective overturn time is relatively short. It is suggested that condensation of CH_4 might produce a stepwise stratification of mean molecular weight that would increase stability, playing the same role as salinity variations in generating layered convection in terrestrial oceans. An alternative suggestion by Flasar (19) is that the lapse rate for $P > 700$ mb is actually stable because of methane condensation that introduces a buoyancy gradient capable of supporting what otherwise would appear to be a superadiabatic (highly unstable) gradient for equilibrium hydrogen. Flasar suggests that the close agreement between the measured lapse rate and the adiabatic gradient for normal hydrogen is just a coincidence.

The He/H_2 ratio is constrained best from the analysis of far-infrared spectra. The collision-induced dipole absorption of $\text{H}_2\text{-H}_2$ and $\text{H}_2\text{-He}$ proves most of the continuum opacity at long wavelengths and is sufficiently well understood theoretically that it is possible to infer the ratio by fitting the thermal IR spectra that are also consistent with temperature profiles determined by Voyager 2 radio occultation measurements. The inferred helium mole fraction (number or volume fraction rather than mass fraction) in the upper atmosphere of Uranus is $15.2\% \pm 3.3\%$ (20). Neptune's is a little higher at $19.0\% \pm 3.2\%$ (21). As expected from formation theories, both are within errors of the solar value of 16%. This contrasts with Saturn, which has only about one-fourth of the solar fraction, presumably due to rainout of He in Saturn's deep interior where it is thought that He becomes insoluble in metallic hydrogen.

Spectroscopic observations indicate that methane (CH_4) is enhanced relative to solar abundance values by a factor of 25–30 in regions where it ought to be well mixed (22). However, because methane condenses to form clouds in both atmospheres, methane abundance above the condensation level can be greatly reduced, variable, and difficult to estimate. Initially, it was thought that the tropopause would act as a cold trap, limiting the stratospheric methane mixing ratio to values no greater than the saturation mixing ratio at the tropopause. Yet Voyager Ultra-Violet Spectrometer data for Neptune imply that stratospheric mixing ratios are at least 10 times this limit (23). This excess, termed oversaturation, suggests stronger vertical mixing on Neptune compared to Uranus, where stratospheric oversaturation is not observed. The large tropospheric enhancement of methane relative to the solar mixing ratio for both Uranus and Neptune is a likely consequence of the planetary formation process. The large fraction of icy materials accreted by these planets should also have resulted in similar enhancements of water and ammonia, even though little has so far been observed in the atmospheres.

Microwave atmospheric observations imply a significant depletion of NH_3 relative to solar values, by a factor of 100–200 in the 150–200 K region of Uranus' atmosphere (24). A possible explanation is that NH_3 is lost to an extensive NH_3 – H_2O solution cloud or that NH_3 is lost to the formation of a cloud of NH_4SH (ammonium hydrosulfide). The depletion might be completely accomplished by formation of an NH_4SH cloud if the $\text{H}_2\text{S}/\text{NH}_3$ ratio is enhanced by a factor of 4 compared to the solar value of 0.2. It seems to require very large enhancement factors of water for the NH_3 – H_2O solution cloud to play a major role in the depletion of NH_3 . The necessary enhancement of H_2S might have resulted from accretion of chondritic meteorites, in which sulfide minerals are found but nitrogen incorporation is not significant. An alternative hypothesis by Lewis and Prinn (25) is that Uranus never acquired much nitrogen because uncondensed N_2 and CO were the dominant chemical forms of N and C in the solar nebula (rather than NH_3 and CH_4) in the region of the terrestrial planets and also, as a result of rapid mixing in the outer parts of the nebula where thermal equilibrium compounds would be NH_3 and CH_4 . So far, there is no direct observation of H_2O or NH_3 on either Uranus or Neptune, nor of H_2S on any Jovian planet except Jupiter. If H_2O is present at solar mixing ratios, condensation clouds might be found at pressures of the order of 100 bars. If enhanced 65 times solar, water would condense at a temperature of 647 K, but would not condense at all beyond that enhancement (24).

Spectral Characteristics of Uranus and Neptune. Methane absorption dominates the visible and near-IR disk-averaged spectra of both Uranus and Neptune (upper part of Fig. 2). At wavelengths below 1 micron, there are numerous methane bands of increasingly greater absorption from the green to the red and near IR. The absorption of red light by methane contributes to the blue colors of Uranus and Neptune. Neptune is bluer than Uranus because of increased absorption in the window regions between the methane bands. This might mean that the visible cloud layer that controls the amount of light reflected backward is thicker and brighter on Uranus, whereas the corresponding cloud layer on Neptune is more transparent and allows more of the light to be absorbed by the deeper atmosphere. Alternatively, the cloud itself might provide the extra absorption needed on Neptune (22); if so, the cloud must absorb even more effectively at wavelengths between 1 and 2.5 microns (29).

The basic disk-averaged characteristics of Neptune and Uranus at near-IR wavelengths are illustrated in the upper right panel of Fig. 2, which displays the ground-based observations of Fink and Larsen (27). The relatively low disk-integrated albedos of both planets are due to the significant amount of methane overlying the visible cloud layer. The albedo peaks that occur in windows of relatively weak absorption, for example, 1.3 and 1.6 μm , are narrower for Uranus than for Neptune, indicating that Uranus has a clearer atmosphere with fewer high altitude hazes to reflect photons that would otherwise be absorbed by underlying methane. In the 1–1.8 μm region methane is the dominant absorber, whereas hydrogen collision-induced absorption (CIA) is the major absorber in the 1.85–2.2 μm range.

The effects of these absorbers on the penetration depths of photons into the atmospheres of these planets are illustrated in the middle panel of Fig. 2. This displays the wavelength-dependent pressure levels at which a unit albedo re-

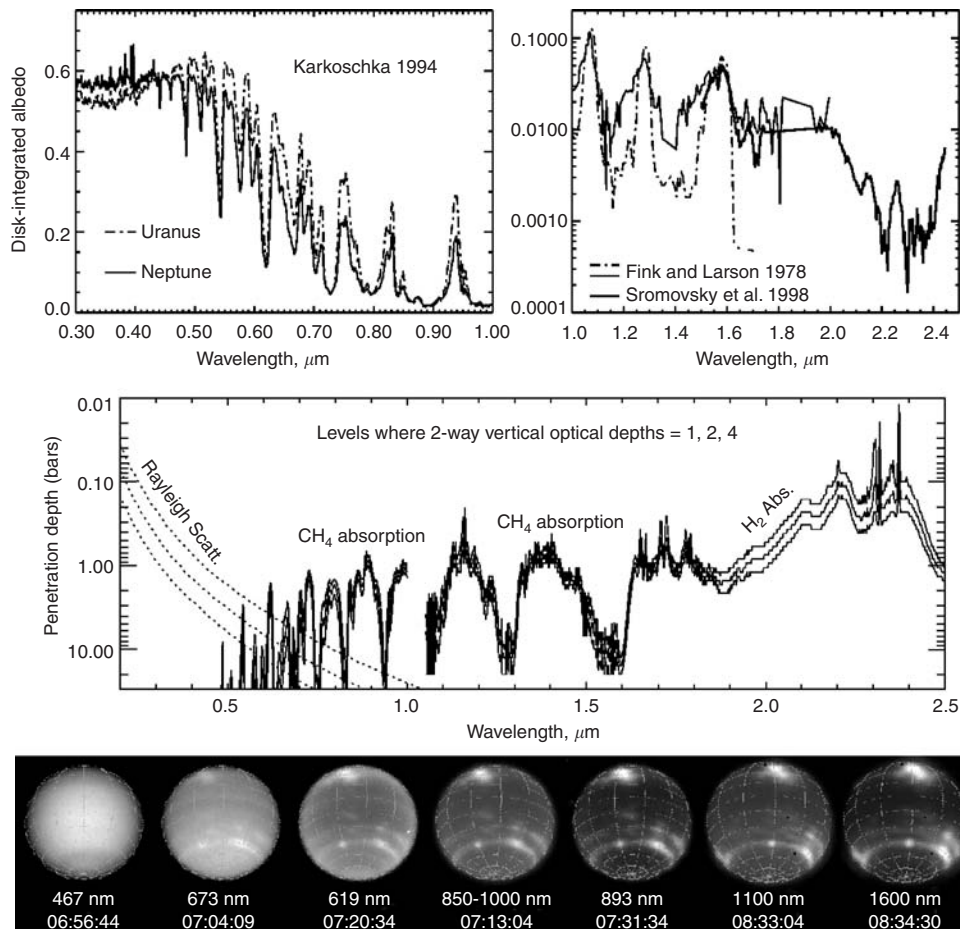


Figure 2. (Top) Reflectivity spectra of Neptune (26–28). (Middle) Penetration depth of photons into Neptune’s atmosphere (30). (Bottom) Appearance of Neptune as a function of wavelength as recorded by HST imaging of WFPC2 and NICMOS (30).

flecting layer would have apparent albedos of 37%, 13.5%, and 1.8% (e^{-1} , e^{-2} , and e^{-4}) (30). This spectrum of sensing depths was computed for Neptune using a tropospheric methane mixing ratio of 0.022 and a stratospheric mixing ratio of 3.5×10^{-4} (22), the saturation mixing ratio is assumed wherever it is less than either of the others. (Penetration depths for Uranus are roughly the same as those for Neptune.) In Fig. 2, we see that the 1.27 μm and 1.59- μm windows probe all the way into the putative H_2S visible cloud deck near 3.2 bars. Were that cloud composed of high-albedo particles, we would expect albedo values near unity, rather than values in the 0.05–0.08 range. On the basis of the near-IR spectrum alone, the cloud is very dark, more transparent, or deeper than indicated by other methods. The exact levels sensed in the strongly absorbing 2.3- μm methane band depend on stratospheric methane mixing ratios, which are not well constrained by current observations. Sample images of Neptune obtained by the Hubble Space Telescope Wide Field/Planetary Camera 2 and Near Infrared

Camera and Multi-Object Spectrometer are shown in the bottom panel of Fig. 2 (30). These illustrate the dominance of Rayleigh scattering at short wavelengths and the effects of methane absorption at long wavelengths.

Neptune has both dark and bright discrete features, but the only discrete features on Uranus seem to be bright features. Dark features on Neptune have the greatest contrast at blue wavelengths (see lower left of Fig. 2), but it is typically a rather small 2–10%. The identity of the blue absorber that creates this contrast is unknown. The contrast between bright clouds and the background atmosphere can also be rather small at short wavelengths, where Rayleigh scattering is important, but it is dramatically enhanced at wavelengths where methane absorption is strong. On Neptune, the contrast can exceed 500:1 at 2 microns (30). The maximum contrast for bright clouds on Uranus is seen at about 1.9 microns and is about 1.8:1 in raw images, but estimated at about 10:1 at high spatial resolution (31). Relatively less contrast is observed on Uranus because its bright cloud features do not reach to pressures as low as those on Neptune.

Temperature Structure. The temperature structures of Uranus and Neptune are very similar, as illustrated in Fig. 3 by dotted and solid curves, respectively. There is remarkably little latitudinal variation in this structure. Given the 98° obliquity (pole inclination relative to orbital normal) of Uranus and Neptune’s 29°, the absence of significant pole-to-pole gradients (for Uranus) or equator to pole gradients (for Neptune) implies some heat transport or compensation effect. The effective emission temperatures on both planets are very low (about 59 K), leading to very long radiative time constants of about 5×10^9 seconds at 400–500 mb (32). This is the ratio of the thermal energy content of an atmosphere to

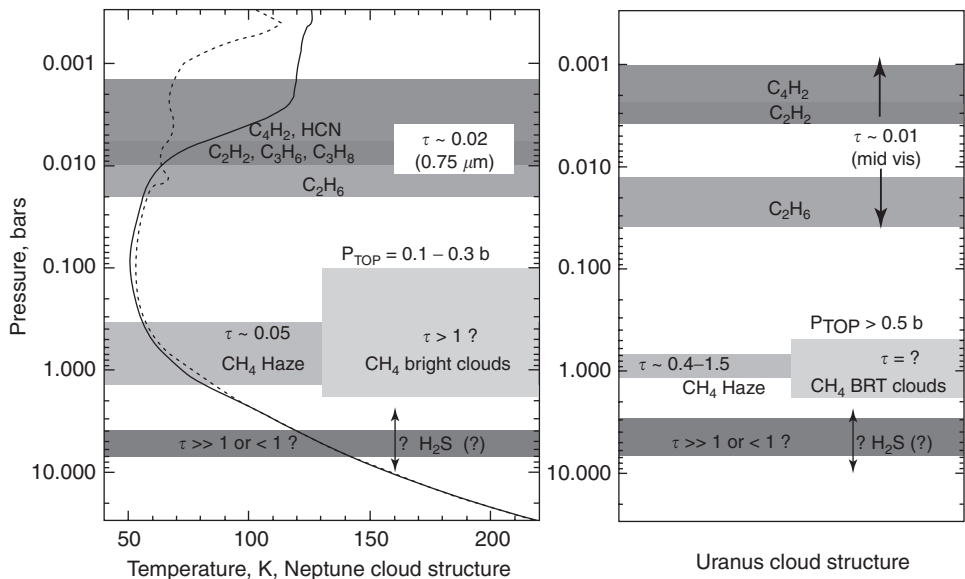


Figure 3. Vertical structure of temperatures and cloud and haze layers. The Neptune temperature profile (solid) differs from the Uranus profile (dotted) mainly in the stratosphere where a larger population of hydrocarbon hazes on Neptune leads to more heat absorption and a warmer stratosphere.

the radiative cooling rate for one local scale height. This radiative time constant is twice as long as Uranus' orbital period. Thus, seasonal variations on Uranus are strongly damped and phase shifted (delayed) by close to 1/4 year. In fact, even though the North Pole (IAU convention) of Uranus had been in darkness for 20 years, the Voyager IRIS instrument found no measurable difference in temperature structure between the two polar regions (33). Voyager 2 arrived close to the Southern Hemisphere solstice, a time when hemispheric thermal contrast should be at a minimum because of the phase lag, and thus it was not able to measure the amplitude of the seasonal response. Because of Uranus' obliquity, the average solar heating for an entire Uranian year is greater at the poles than at the equator. However, seasonal model predictions of a somewhat cooler equator were not confirmed by Voyager observations. The main latitudinal variation in temperature on both Uranus and Neptune is consistent with the thermal wind equation. This equation is a proportionality between vertical wind shear and horizontal temperature gradients, which is valid for vertical hydrostatic balance (between pressure and gravity) and horizontal balance between coriolis forces and horizontal pressure gradients. The sign of the derived vertical wind shear indicates that the zonal winds decay as height increases (discussed later).

Cloud Structure. Our current limited understanding of cloud structures on Uranus and Neptune is illustrated in Fig. 3. The saturation vapor pressure curve for a CH_4 mixing ratio of 2% suggests a methane cloud base near 1.4 bars on Neptune and 1.2 bars on Uranus, although nucleation at somewhat higher altitudes is likely because it would probably require some degree of supersaturation (34). Radio occultation observations of refractivity gradients in this region agree with the 1.2 bars on Uranus (35) but suggest a cloud base near 1.9 bars on Neptune (36). Neptune's stratospheric hazes (34,37,38), which significantly reduce Neptune's shortwave albedo, have a relatively low total equatorial optical depth of ~ 0.02 at $0.75\ \mu\text{m}$ (39). Larger optical depths found for the global average might be due to contributions from unresolved high-altitude, isolated, bright methane clouds. The optical depth of the methane cloud at red wavelengths is relatively small; on Neptune, it ranges from about 0.05 near the equator to 0.3 near 25°S (22), and on Uranus from 0.4 (40) to 1.3–1.5 (41,42); there is evidence for lower values (several tenths) for central disk observations, and higher values are inferred from disk-integrated values (43). Much larger opacities are possible for discrete features. The presence of an H_2S cloud is inferred from microwave spectra that probe the deep atmosphere; enhancement by a factor of 30 relative to the solar mixing ratio is not directly measured but is inferred to explain the very low abundance of NH_3 (44,45). In this scenario, the formation of a very deep NH_4SH cloud consumes the excess NH_3 . When the corresponding H_2S condensation curve is computed, we find that it intersects the temperature profile in the 6–9 bar region, above which we might expect to see a cloud of H_2S . The presence of an opaque cloud at pressures deeper than 3.6 to 3.8 bars is inferred from hydrogen quadrupole line widths (40). However, attempts to find direct evidence of H_2S in optical spectra have failed (46). Thus, we cannot completely rule out the possibility that both NH_3 and H_2S are very depleted, as advocated by Romani et al. (37) and that the visible cloud deck is a thin ammonia ice cloud or that no cloud at all is present in the 3–10 bar region, though that would conflict with the quadrupole results. An ammonia cloud, if present, might

also form near the 7-bar level. Current models also have difficulty explaining thermal infrared observations of Neptune, which seem to require that the methane cloud be much more opaque at a significantly higher altitude or horizontally heterogeneous (21).

Uranus is clearer than Neptune, perhaps because it lacks internal heat to drive the mixing needed to keep particles suspended for long times. The global average methane haze layer on Neptune is relatively thin (~ 0.1), but the optical depths of discrete bright cloud features are probably much greater. On Neptune, these features reach 50–100 km above the mean cloud level, reaching to pressures of 100–200 mbar. On Uranus, discrete bright clouds do not rise much above 500 mb (31).

Horizontal cloud structures for Uranus and Neptune are illustrated by sample Voyager and HST images in Figs. 4–7. Voyager imaging of Uranus (Fig. 4) in 1986 displayed a rather bland appearance. An approximate true-color image (left image in Fig. 4) displays neither banding nor discrete cloud features. The extremely enhanced false-color image in the middle of Fig. 4 shows that there was a latitudinal variation in the amount of UV-absorbing haze in the atmosphere; the greatest absorption occurred near the visible pole of Uranus (IAU south) where it appears as relatively orange. Discrete cloud features are illustrated in the time sequence of images at the right of Fig. 4, which were made using Voyager's orange filter.

Hubble Space Telescope images of Uranus made in 1997, 11 years after the Voyager encounter, reveal the first discrete cloud features in its Northern Hemisphere. One feature is barely visible near the right-hand limb in the upper left

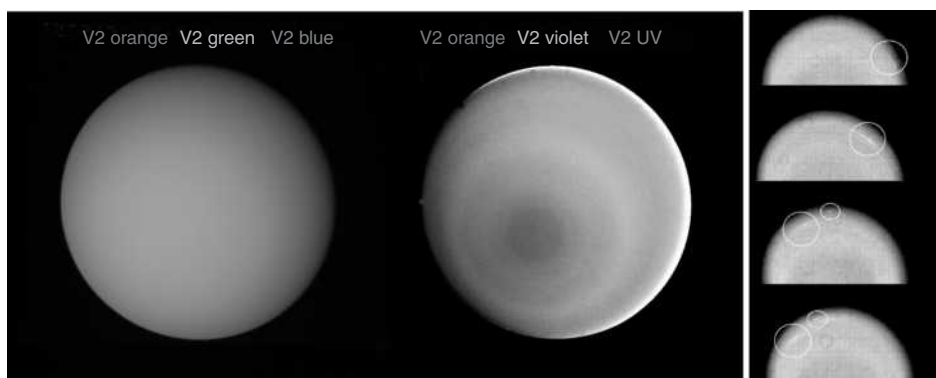


Figure 4. [JPL P29478]. Voyager 2 images of Uranus taken on 17 January 1986 using blue, green, and orange filters to make the true-color composite (left), which displays a virtually featureless disk. UV, violet, and orange filtered images were shown as blue, green, and red components in the extremely enhanced false-color image (middle), which reveals polar bands of UV-absorbing haze particles, centered on the South Pole of Uranus (IAU convention). Even in this view, no discrete cloud features are apparent. [Voyager 2 JPL P29467.] The right-hand time sequence of orange-filtered images from 14 January 1986 shows the motion of two small bright streaky clouds that were the first discrete features ever seen on Uranus. Uranus is rotating counterclockwise in this view, as are the clouds, though more slowly than Uranus' interior, revealing that low latitude winds on Uranus are retrograde, as are the winds on Neptune. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

image, which was made with a 619-nm filter (the wavelength of a weak methane band). Much better contrast between discrete cloud features and the background atmosphere was obtained at near-IR wavelengths using the HST NICMOS camera, as illustrated in the middle and right-hand images of Fig. 5.

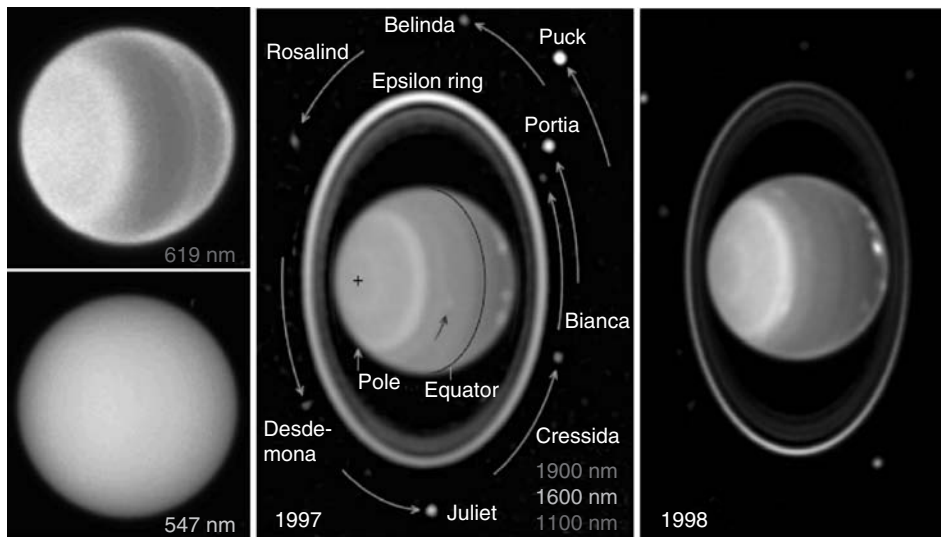


Figure 5. (Left) [STScI PRC97-36b, NASA and H. Hammel]. HST WFPC2 images of Uranus on 31 July and 1 August 1997. Although little contrast is seen at 547 nm (blue), a banded structure and the first discrete Northern Hemisphere cloud are visible at 619 nm (upper image, colored red). (Middle) [Space Telescope Science Institute STSCI-PRC97-36A and E. Karkoschka]. This false-color 1997 image is a composite of near-IR images taken by the Near Infrared Camera and Multi-Object Spectrometer (NICMOS) at wavelengths of 1.1 μm (shown as blue), 1.6 μm (shown as green), and 1.9 μm (shown as red). Absorption by methane gas limits the depth at which reflected sunlight can still be seen at 1.1 and 1.6 μm , and absorption by hydrogen is most significant at 1.9 μm . The blue exposure probes atmospheric levels down to a few bars, responding to scattering by aerosols and by atmospheric molecules above this level. The green component is least sensitive to methane absorption, sensing down to 10 bars or more, but sees much less Rayleigh scattering per bar than the blue component, so that a dark absorbing cloud near 3 bars would result in more blue than green in regions that were clear above the cloud, perhaps accounting for the blue color at midlatitudes. The red component can only sense down to about the 2-bar level, and sees the least contribution from Rayleigh scattering, so that very little red is seen in regions that are clear to 3 bars. The green color around the South Pole suggests significant local haze opacity at pressures near 2–3 bars. The red color of the discrete features near the northern (right) limb indicate relatively high-altitude clouds that reflect sunlight before much absorption has taken place. The curved arcs in the central image indicate motions in 90 minutes of cloud features and eight of the 10 small satellites discovered by Voyager 2. The area outside the rings was enhanced to make the satellites more visible. The images also show the bright epsilon ring, which is wider and brighter in the upper part of the image, and two fainter inner rings. (Right) [STScI-PR98-35, NASA and E. Karkoschka]. This Hubble Space Telescope near-IR image of Uranus on 8 August 1998 shows a number of new cloud features in both hemispheres. The false-color image was created in a manner similar to the first, except that the rings and satellites were not separately brightened. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

Voyager 2 images of Neptune in 1989 (Fig. 6) provided a rich bounty of detail concerning horizontal cloud structure and revealed many discrete bright and dark cloud features, as well as bright and dark bands. Many of these structures were unique to Neptune and have not been seen before or since. HST imaging of Neptune beginning in 1994 showed some cloud banding similar to the 1989 Voyager images, but many changes were found, including the disappearance of Neptune's Great Dark Spot (discussed later).

Weather Phenomena on Neptune

Dark Spots. Voyager identified two prominent dark oval features in Neptune's Southern Hemisphere; both are thought to be manifestations of anticyclonic eddies, as illustrated in Figs. 6 and 7. (Anticyclones rotate counter-clockwise in the Southern Hemisphere.) The Great Dark Spot (GDS) was about an Earth diameter long. It was seen for about 8 months during Voyager's approach to Neptune but has not been seen since. It exhibited several unusual dynamic features; some of them are illustrated in the middle panel of Fig. 6. Its

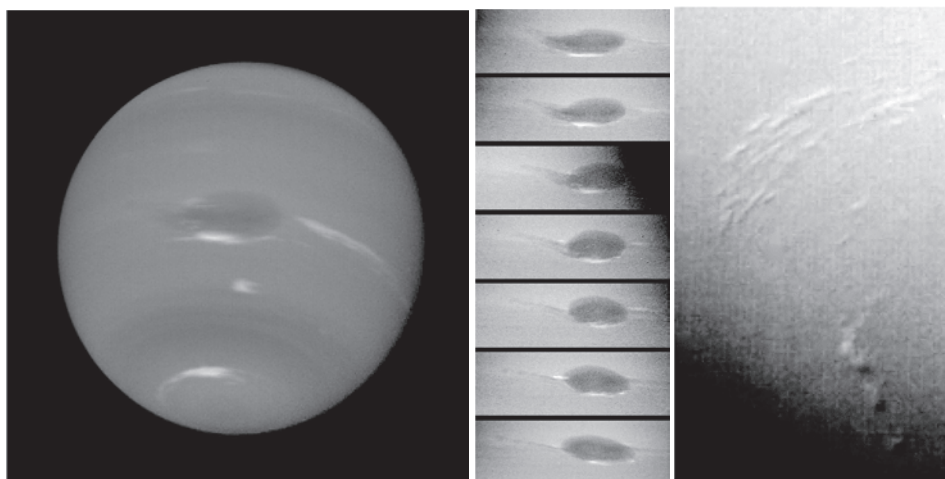


Figure 6. (Left) [JPL P34606]. Voyager 2 image of Neptune. Color composite formed from green, blue, and red filtered images taken on 18 August 1989. South is down, and the South Pole is tipped toward Earth. The Great Dark Spot (GDS) seen at the center of the image is about 13,000 km (about the diameter of Earth) by 6,600 km, and at a latitude of 18°S at the time of this image. The bright cloud to the south of the GDS, termed the companion, is at relatively high altitude compared to the blue features, and was very prominent in ground-based images made at methane-band wavelengths (51). The bright clouds at the edge of the dark circumpolar band, called south polar features (SPF), are at a latitude of about 67°S; these are highly variable on short timescales and exhibit vertical relief of about 75–150 km, as inferred from shadows seen in Voyager images. (Right) [JPL P34668]. This image of Neptune's south polar regions near 68°S, made on 23 August 1989, shows the first cloud shadows ever recorded by Voyager on any planet (the Sun is to the left). (Middle) [JPL P34610]. This time sequence of remapped images of Neptune's GDS, from top to bottom, provides views at intervals of about 18 hours, revealing its strange wobble and shape changes, which occur in a period of 193 hours. Its mean dimensions were approximately 38° in longitude and 14° in latitude and had modulation amplitudes of 7.4 and 1.5°, respectively (47). It also drifted toward the equator at a rate of about 1° per month. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

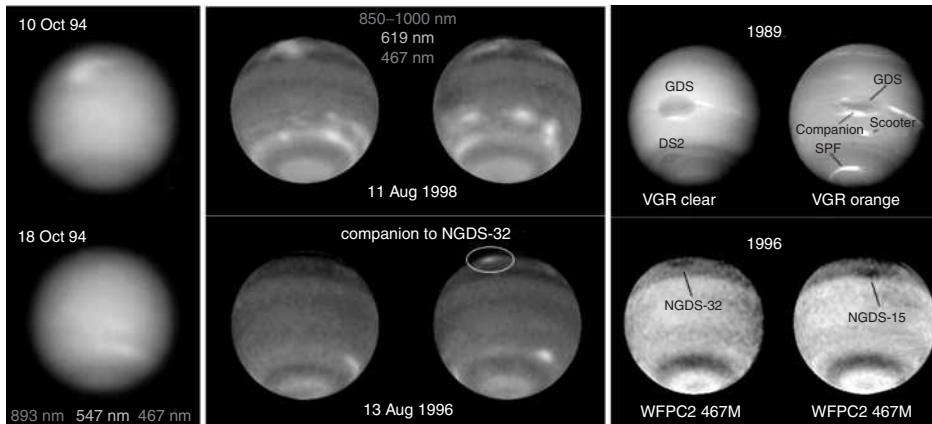


Figure 7. (Left) [Space Telescope Science Institute PRC-98-34 and H. Hammel]. HST composite color image made from images at blue, orange, and near-IR wavelengths and corrected for limb darkening. (Middle) [STScI and L. Sromovsky]. This quartet of false-color images made using HST shows persistent banded structure and an increase in the number of bright clouds between 1996 and 1998. (Right) [STScI and L. Sromovsky]. Views of dark spots seen by Voyager (GDS in top row) and HST (NGDS-32 and NGDS-15 in bottom row). The HST images (48) are composites of several images taken at slightly different times. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

shape and orientation varied cyclically in a period of 193 hours (47). The longitudinal width varied from 31 to 45° , and its latitudinal extent varied from 12.5 to 15.5° . This behavior was successfully reproduced using a Kiva vortex model (49). This model predicts anticyclonic circulation (counterclockwise in the Southern Hemisphere), although no direct measurements have been able to test this prediction. The GDS also drifted toward the equator at a steady rate of 1.24° /month (47); this would have placed it on the equator in November 1990, although theoretical models imply that it would have dissipated before that point (50). This behavior is in marked contrast to that of Jupiter's Great Red Spot, which has remained at essentially a fixed latitude for centuries. The GDS was accompanied by a prominent bright companion cloud positioned at its southern boundary (Figs. 6 and 7). In ground-based methane band imaging at the time of the Voyager encounter, the companion was the brightest feature on the planet (51) and seems to be similar to orographic clouds. It is likely caused by vertical deflection of flow around the GDS, producing methane condensation during the upward part of the flow and evaporation on the downward return leg (52). The first Hubble Space Telescope images that had sufficient image quality to show the subtle contrast of the GDS at blue wavelengths and sufficient longitudinal coverage were made in 1994, after the Hubble repair mission. The GDS was not seen at that time, and no other southern dark spot has been seen since.

The second prominent dark spot observed by Voyager (DS2 in Fig. 7) was much smaller than the GDS (about 20° in longitude by 6° in latitude), and had several unique dynamic characteristics of its own (47). Its latitudinal position oscillated at an amplitude of 2.4° about its mean latitude of 52.5° S in a period of

865 hours. During this oscillation, its zonal speed increased in step with the zonal wind shear, resulting in a 48° longitudinal oscillation amplitude. Bright clouds formed near the center of the dark spot during the northern half of the oscillation and declined during the southern half. Strangest of all, its mean drift rate relative to Neptune's interior matched to great precision the mean drift rate of the south polar features (discussed later); the two mean positions remained on opposite sides of Neptune and at mean latitudes differing by 20° . The DS2 feature has not been seen since the Voyager encounter.

Two new dark spots were discovered in HST images (Fig. 7); both were in the Northern Hemisphere. NGDS-32 was discovered first in 1994 images (53) at about 32° N latitude, and a bright companion cloud was seen at its equatorward edge. This dark spot is comparable in size to the GDS but is harder to observe because the latitude circle where it is found is close to the northern limb. NGDS-15 was discovered first in 1996 HST images (54) close to latitude 15° N. It is somewhat smaller than NGDS-32 and spans about 20° in longitude and 11° in latitude. It is unusual in having no bright companion cloud. NGDS-15 is also unusual in being close to the latitude at which modeled dark spots tend to dissipate rapidly. Both NGDS-32 and NGDS-15 differ from the Voyager GDS by exhibiting no measurable latitudinal drift, a characteristic that is baffling to modelers because model dark spots always drift equatorward on Neptune. The modeled drift rate depends on the latitudinal gradient in the zonal wind speed. If that gradient is locally perturbed in the vicinity of the dark spot latitudes, it may be possible to get consistency between model and measured behavior.

South Polar Features. Rapidly varying small bright features clustered near $62\text{--}70^\circ$ latitude, termed south polar features, were the first clouds Voyager observed that had shadows (see Fig. 6). These features were observed almost exclusively in a region less than 180° wide, and though the mean position where they were seen drifted only slightly relative to the interior, the individual cloud elements actually moved relatively quickly at a prograde rate of about 200 m/s. Their lifetimes were so short that they could not be tracked during a full rotation of Neptune. At low resolution, the features blend together so that they sometimes appear to form a plume (the southernmost prominent bright cloud in the left image of Fig. 6). But in high-resolution images, they can be seen as small individual elements (see the right image of Fig. 6). The vertical relief implied by shadows is 50–150 km (51).

Waves and Bands. Cloud patterns on Neptune suggest the existence of atmospheric wave motions. Voyager imaging showed that near 20° S latitude, there was a clustering of bright cloud features at two regions spaced approximately 180° apart in longitude; one region was in the vicinity of the GDS. This suggests a wave that has two complete oscillations within 360° of longitude (wave number 2). Neither the pattern nor the GDS have been seen since the Voyager encounter. The phase-locked motions of the region of SPF formation and DS2 suggest a similar wave interaction. A more obvious wave example is the dark band between 55° S and 65° S, visible in blue-filtered images (right group in Fig. 7). When viewed in polar projection, the band appears as a circle offset from Neptune's South Pole, leading to a 2° sinusoidal modulation in latitude of the wave boundaries as a function of longitude. The wave can be seen as a difference in tilts of the band in the two images shown for August 1998 in the middle group

of Fig. 7 where the band appears greenish because of the false-color scheme. Although DS2, which was nestled in one of the northern excursions of the wave in 1989, is no longer visible (at least in HST images), the wave structure appears to have persisted from 1989 through at least 1998.

Secular Variations on Uranus and Neptune. In recent years, Uranus has revealed greater weather activity as its Northern Hemisphere continued its emergence out of decades of darkness. First seen in HST images of Uranus (31,55), bright Northern Hemisphere clouds have now become visible from the ground (56). Over a longer time period, Uranus has exhibited rather strong disk-averaged albedo variations, including a 14% increase from 1963 to 1981 (during which the sub-Earth latitude varied from near zero to 68°S) (57). Additional ground-based observations of a peak blue reflectivity in 1985 and a 7% decline from 1985 to 2001 are reported by Karkoschka (58), who also showed that there is a hemispheric asymmetry in brightness; the Northern Hemisphere is darker, and much of the recent decline in brightness has to do with geometric effects as more of the darker Northern Hemisphere comes into view. He also showed that additional physical effects played a role in earlier brightness increases and suggested that significant physical effects would also appear in the near future.

Recently, Neptune also exhibited a relatively steady brightening at visible wavelengths of 1.4–1.9% from 1996 to 1998 (30) and almost 10% from 1990 to 2000 (59); this dramatically breaks from the previous inverse correlation with solar UV variations. Neptune's atmosphere also exhibits dynamic activity during decade-long timescales, as established by ground-based imaging of changing distributions of bright features and ground-based photometry of varying light curves. During 1976 and 1987, it appears that much brighter cloud features were present than have been seen during or since Voyager (60). Ground-based observations of bright cloud features at latitudes where none were seen by Voyager, large changes in light curve amplitudes from year to year (61), and factor of 10 changes in near-IR brightness (62) also suggest major developments or large latitudinal excursions during a several year period. Sromovsky et al. (29) showed that the spectral character of the 1977 "outburst" could be matched by a factor of 7 increase relative to August 1996 in the fractional area covered by high-altitude, bright, cloud features.

Zonal Mean Circulation and Its Stability. Uranus and Neptune provide strong evidence that rotation dominates solar radiation and internal heat flux in determining the form of a planet's zonal circulation. Given that Uranus has had one pole in sunlight for ~ 20 years, it would have been natural to expect strong meridional circulation between the two hemispheres. Yet, the observed temperature difference between Uranus' long dark hemisphere and that heated by the Sun is very small and opposite in sign to this expectation. Part of the explanation was already discussed in the section on thermal structure, that is, the long radiative time constant removes most of the interhemispheric thermal contrast. The form of the circulation turned out to be zonal (parallel to lines of constant latitude).

Voyager's inability to find many cloud features on Uranus in 1989 hampered efforts to characterize fully its mean zonal circulation. Yet, even with sparse sampling, it was clear that the circulation was retrograde near the equator (atmospheric parcels fell behind the planet's rotation) and prograde at high

latitudes. Radio occultation measurements provided one low latitude measurement (Fig. 8) that was critical in confirming this picture. Uranus' cloud features have remained difficult to observe since that time, until recently. Improved near-IR imaging capabilities (NICMOS and Keck adaptive optics imaging) and the emergence of the Northern (IAU convention) Hemisphere into daylight have brought many new cloud features into view. New measurements, summarized by Hammel et al. (53) and shown in Fig. 8, suggest a possible small difference between hemispheres. To first order, however, the new data are consistent with the Voyager results and suggest a relatively stable and symmetrical circulation that bears a strong resemblance to Neptune's in form but is considerably weaker in amplitude.

The most detailed definition of Neptune's zonal mean circulation is provided by Voyager 2 imaging observations (64). A strong retrograde equatorial jet at 400 m/s (895 mph) and a weaker prograde jet at 250 m/s are the main features of the circulation (Fig. 8). The prograde jet on Uranus moves at about 200 m/s and at a lower latitude (60° vs. $70\text{--}75^\circ$ for Neptune). The Uranian equatorial jet is much weaker than Neptune's (about 100 m/s retrograde vs. 400 m/s for Neptune) and covers a narrower latitude range ($\pm 25^\circ$ vs. $\pm 50^\circ$ for Neptune).

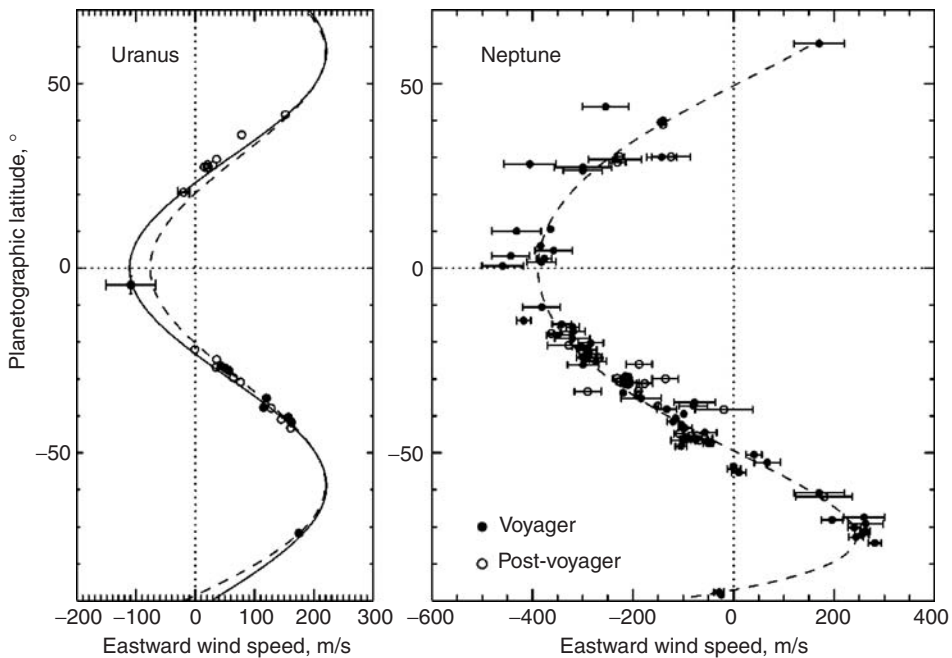


Figure 8. Zonal mean circulations of Uranus and Neptune. Voyager results (filled circles) and post-voyager results (open circles) are roughly consistent with an unchanging symmetrical circulation for both planets, although somewhat better fits to the Uranus observations are obtained with slightly asymmetrical profiles. Data points are from the Uranus compilation by Hammel et al. (63), and the Neptune Voyager observations of Limaye and Sromovsky (64) and HST observations by Sromovsky et al. (30). The post-Voyager results for Uranus are a combination of HST and Keck imaging from 1997–2000, the post-Voyager Neptune results are entirely from HST imaging from 1994–1998.

From the wind measurements displayed in Fig. 8, it seems that there is much less variability in the measurements of Uranus. This is in part the result of very rapid evolution of cloud features on Neptune, which reduces the typical time interval over which a cloud feature can be tracked and thus reduces the accuracy in speed that can be achieved. It is also true that part of the variability in the Neptune measurements comes from true eddy motions. The cloud features tracked on Uranus, on the other hand, seem to be more stable, allowing longer observation periods. This might be so because the cloud features are less numerous and less visible on Uranus, so that only the more intense and long-lasting features can be seen at all. It is also likely that due to far lower internal energy flux on Uranus, there may be less eddy activity to start with, as indicated by the small number of high altitude clouds and the relatively clear upper troposphere.

The equatorial retrograde winds on both planets might be a consequence of angular momentum conservation in the presence of an axisymmetric meridional circulation. Because distance to the spin axis decreases with latitude, so does the angular momentum of atmospheric parcels at rest with respect to the rotating planet. But if atmospheric mass at midlatitudes were to move equatorward, it would tend to slow its rotational speed to conserve its angular momentum, leading to a retrograde wind near the equator at the level of the equatorward flow. Midlatitude gas parcels moving poleward would tend to speed up as they get closer to the spin axis, leading to a prograde circulation at high latitudes. The reverse flow at return levels would tend to produce the opposite effects. This meridional flow may be consistent with the observed zonal circulations, but there are no direct observations that could confirm or deny the existence of the meridional circulation. Such a model would not explain the equatorial prograde jets of Jupiter and Saturn that clearly require transport of angular momentum by eddies.

The solar irradiance at Uranus is about 3.8 W/m^2 , and the planet-wide average is about 0.65 W/m^2 in absorbed flux. Neptune, on the other hand, is exposed to a solar irradiance of only 1.5 W/m^2 and absorbs an average of about 0.26 W/m^2 compared to its internal heat flux of 0.4 W/m^2 (65). Thus, the two planets actually have about the same total heat flux available for generating atmospheric motions. Their different atmospheric motions might be due to differences in the latitudinal distributions of the fluxes. Based on its modest spin axis inclination, Neptune's equator will receive much more solar heat than its poles, even when averaged for a year, whereas the 98° inclination of Uranus results in its poles receiving 50% greater solar input than the equator. Based on calculations for Saturn, which has almost the same obliquity as Neptune, latitudes poleward of 60° should receive only half of the average solar flux received near the equator. Despite these different heat input distributions, there is little latitudinal variation in atmospheric temperatures on either planet, as noted in the discussion of thermal structure. To equilibrate the temperatures across all latitudes would require different magnitudes of horizontal heat transport and (even different directions) on Uranus and Neptune. This may account for some differences in circulation, but the point made by Ingersoll (66) is worth noting: there is not much connection between energy sources and speeds and patterns of outer planet circulations. Jupiter has 20 times the available power to drive circulations, yet it has only one-third the wind speeds. Although baroclinic eddies might

provide the needed horizontal heat transport on Uranus, Friedson and Ingersoll (67) suggest that the internal heat flux on Neptune might reduce latitudinal gradients, acting as a sort of thermostatic control, efficiently providing extra heat in regions that are slightly cooler and reducing heat in regions that are slightly warmer. For this mechanism to work, the solar energy absorption must occur in regions within or close to the free convection zone that extends into the interior. But how can the high winds of Uranus and Neptune and Neptune's highly variable weather phenomena be maintained? The answer seems to be that these atmospheres have very low dissipation and thus take very little power to keep them running, much like a well-lubricated ball bearing.

We don't know very much about the winds below or above the level of the visible cloud features. We can use Voyager 2 measurements of horizontal temperature gradients to estimate the vertical wind shear by using the thermal wind equation discussed earlier. We find that the winds on both Uranus and Neptune decay as height increases, suggesting frictional dissipation in the stratosphere (33,66). The vertical shear is relatively low and requires about 10 scale heights (the pressure drops by a factor of $1/e$ for each scale height) to damp to zero velocity. To what depth below the clouds the winds continue to increase is unknown. We do know from studies of Neptune's gravity field (discussed earlier) that its winds cannot maintain the same speed throughout the interior but must damp to low values within a small fraction of the planet's radius.

The Voyager measurements of Neptune's zonal circulation (Fig. 8) are at least roughly consistent with earlier ground-based and subsequent HST observations (53,54). However, it is not clear that this circulation is stable in detail, or whether there is more latitudinal variation than Voyager observations have indicated. According to the theory of LeBeau and Dowling (50), the detailed curvature of the zonal wind latitudinal profile is important because it determines the drift rate and lifetime of Great Dark Spots. Thus, measurements of both zonal wind and discrete feature latitudinal drifts put strong constraints on such theories. Recent HST observations (30) are beginning to indicate a consistent pattern of small deviations from the Voyager profile.

Satellite Systems

Uranus and Neptune, like Jupiter and Saturn, have systems of regular satellites in prograde orbits that lie near the planets' equatorial planes. The two planets also have irregular satellites. Neptune's moon Triton is the most significant. Because of their orbital similarities and prograde orbits, regular satellites, it is thought, formed with the planet in a common process, rather than being captured after the planet's formation. During the early stages of giant planet formation, the local environment becomes hot enough to vaporize the constituents that later cool to form the solids, which subsequently accrete into satellites. The thermal history of each planet and the timing relative to the blowing out of nebular gas by the intense solar wind generated during the Sun's T Tauri phase are thought to determine the characteristics of the satellite systems.

Satellites of Uranus. Uranus has 20 known satellites; five were known prior to Voyager's 1986 encounter with the Uranian system (Fig. 9), and 11 were

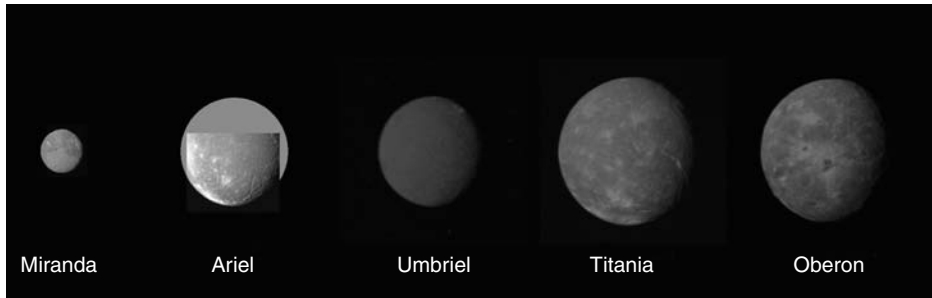


Figure 9. [JPL P30054]. Voyager 2 montage of the five largest satellites of Uranus in order of increasing distance from Uranus (Miranda, Ariel, Umbriel, Titania, Oberon) and correct relative sizes and brightness. Similar to Fig. 15. p. 52, *Science* 233. Imaging Team Report. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

discovered in Voyager images (Table 3). The names of the Uranian satellites are derived from the writings of Shakespeare and Pope. The four largest regular satellites (Ariel, Umbriel, Titania, and Oberon) have low inclination and low eccentricity. Most formation models require that the satellites formed after the

Table 3. Satellites of Uranus

Satellite	Orbital radius (km)	Radius (km) ^a	Mass (10 ²¹ g)	Sidereal period (days) ^b	Density ^a	Orbital inclina- tion, deg	Geomet- ric al- bedo ^a	Discovery date ^c
Cordelia	49,750	20	?	0.335	1.3	0.08	0.07	1986a
Ophelia	53,760	21	?	0.376	1.3	0.10	0.07	1986a
Bianca	59,160	26	?	0.435	1.3	0.19	0.07	1986a
Cressida	61,770	40	?	0.464	1.3	0.01	0.07	1986a
Desdemona	62,660	32	?	0.474	1.3	0.11	0.07	1986a
Juliet	64,360	47	?	0.493	1.3	0.07	0.07	1986a
Portia	66,100	68	?	0.513	1.3	0.06	0.07	1986a
Rosalind	69,930	37	?	0.558	1.3	0.28	0.07	1986a
Belinda	75,260	40	?	0.624	1.3	0.03	0.07	1986a
1986U10	76,000	40	?					1999b
Puck	86,000	77	?	0.762	1.3	0.32	0.07	1985a
Miranda	129,780	236	0.063	1.414	1.21	4.22	0.32	1948c
Ariel	191,240	579	1.27	2.520	1.67	0.31	0.39	1851d
Umbriel	265,970	585	1.27	4.144	1.40	0.36	0.21	1851d
Titania	435,840	789	3.49	8.706	1.72	0.10	0.27	1787e
Oberon	582,600	761	3.03	13.463	1.63	0.10	0.23	1787e
Caliban	7,169,000	49	?		1.5	139.68	0.07	1997f
Stephano	7,948,000	10	?		1.5		0.07	1999f
Sycorax	12,213,000	95	?		1.5	152.67	0.07	1997g
Prospero	16,568,000	15	?		1.5		0.07	1999h
Setebos	17,681,000	15	?		1.5		0.07	1999i

^aRef. 68: From ssd.jpl compilation on 24 April 2002.

^bRef. 69.

^cDiscoveries are dated by data acquisition date (a = Voyager 2, b = Karkoschka, c = Kuiper, d = Lassell, e = Herschel, f = Gladman, g = Nicholson, h = Holman, i = Kayelaars).

impact event that tilted Uranus' spin axis and that they evolved to a state of synchronous rotation in which the satellite always keeps the same side facing the planet. Synchronous rotation has been approximately confirmed for all five of the largest satellites by Voyager 2 imaging (70). These satellites are generally denser and darker than the moons around Saturn and have a rocky fraction of 50% or more. One theory is that a giant impact generated a disk that was ice-poor because the shock energy converted methane and ammonia into largely uncondensable CO and N₂. An alternative theory is that the nebular gas in the vicinity of Uranus was more primitive and unprocessed, so that C and N were more tied up in CO and N₂ and less in CH₄ and NH₃. The impact theory has the advantage of possibly generating amorphous solid carbon, which might help to explain the relatively dark color of the satellites. Oriel and Umbriel have a dense population of large impact craters; the population is especially dense for diameters of 50 to 100 km, similar to that observed for many of the oldest and most heavily cratered objects in the solar system. Titania and Ariel have very different crater populations; there are far fewer large craters, and numerical frequency increases in smaller crater sizes, indicative of a secondary impact population and younger surfaces. Miranda's crater distribution looks like that of Oberon and Umbriel but has an average factor of 3 greater number of any given size (70).

Umbriel and Ariel have similar size and mass but dramatically different surface characteristics, indicating differences in evolution or composition. Umbriel (see enlarged view in Fig. 10) is by far the darker, displays a weaker spectral signature for water ice, exhibits very little albedo contrast across most of its surface, and shows no evidence of crater rays at Voyager resolution. The

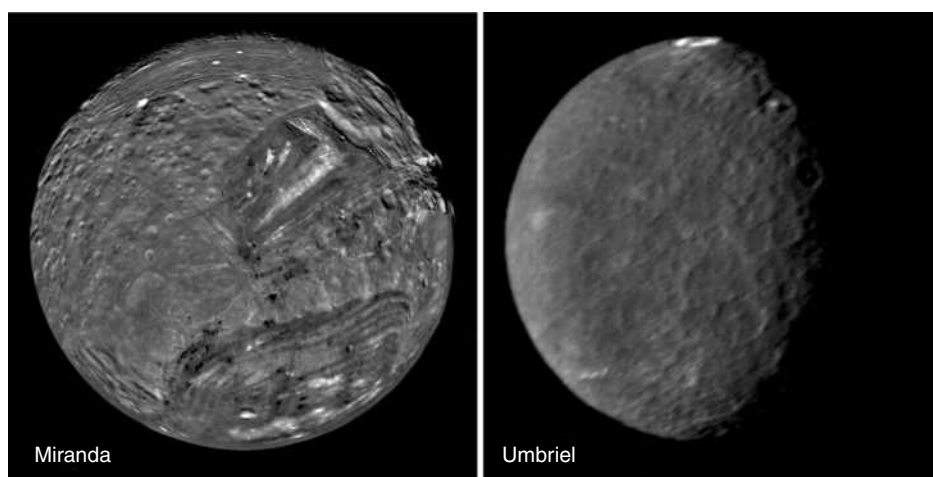


Figure 10. (Left) [USGS P30230]. South polar view of Miranda, produced by a mosaic of nine images obtained by Voyager 2 on 24 January 1986. Older, heavily cratered terrain appears to have low albedo contrast, whereas the younger complex terrain is marked by bright and dark bands, scarps, and ridges. (Right) [JPL P29521]. Southern Hemisphere of Umbriel imaged by Voyager 2 on 24 January 1986. This darkest of the five large moons of Uranus has an ancient, heavily cratered surface and little albedo contrast except for the strange bright ring near the top of the image, which may be a frost deposit associated with an impact crater.

global albedo pattern suggests a young fresh surface, but the crater population suggests a very ancient surface. The surface of Ariel, on the other hand, is much brighter and seems to be geologically younger. Old, large population I craters have been lost, presumably by some combination of viscous relaxation, as indicated by the slumped configuration of its largest existing crater, and extrusion of material over the surface (indicated by smooth plains). Flows on Ariel (and also Titania) are more likely to be a mixture of ammonia and water, methane, or CO clathrates, due to the lower melting points of these mixtures.

Miranda (see enlarged view in Fig. 10) is the smallest of Uranus' large satellites, the closest to Uranus, and is thus most affected by gravitational focusing of external impactors. Its surface is composed of two very different terrain types: an old, heavily cratered terrain without much albedo contrast and a young, complex terrain that has scarps, ridges, and bright and dark bands.

The 11 inner satellites are small and very dark (Table 3). The increasing ice content (inferred from lower density) of satellites closer to Uranus might be due to the higher temperatures closer to Uranus that promote the conversion of CO and N₂ to CH₄ and NH₃. Most of the satellites have nearly circular orbits close to the equatorial plane of Uranus, but the outer four are much more elliptical. Additional small satellites are probably present near the rings and control the narrow rings.

Small Satellites of Neptune. Neptune has eight known satellites (Table 4). Voyager was able to resolve surface features clearly on three of the four largest: Triton, Proteus, and Larissa (Fig. 11), but never got close enough to Nereid to obtain a detailed image. The six inner satellites, all discovered by Voyager in 1989, range in size from 29 to 208 km in radius. They all have low geometric albedos of 0.06–0.08 at visible wavelengths and are gray in color. These are darker than Nereid, which has an albedo of 0.155 (71). The inner satellites are in nearly circular orbits within five planetary radii, whereas Nereid is on a distant and very eccentric orbit that extends from 57 to 385 Neptune radii, suggesting

Table 4. **Satellites of Neptune**

Satellite	Mean orbital radius, km ^b	Radius, km ^a	Mass, 10 ²¹ g ^a	Sidereal period, days ^c	Density, g/cm ³ ^a	Orbital inclination, deg	Geometric albedo ^a	Discovery date ^d
Naiad	48,227	29	0.13	0.296	1.3	4.7	0.060	1989a
Thalassa	50,075	40	0.34	0.312	1.3	0.2	0.060	1989a
Despina	52,526	74	2.25	0.333	1.3	0.1	0.059	1989a
Galatea	61,953	79	2.7	0.429	1.3	0.1	0.063	1989a
Larissa	73,548	96	4.8	0.554	1.3	0.2	0.056	1989a
Proteus	117,647	208	49	1.121	1.3	0.6	0.061	1989a
Triton	354,760	1,353	21,400	5.877	2.07	156.8	0.756	1846b
Nereid	5,513,400	170	31	360.16	1.5	27.6	0.155	1949c

^aRef. 68.

^bRef. 3.

^cRef. 69.

^dDiscoveries are dated by data acquisition date (a = Voyager2, b = Lassell, c = Kuiper).

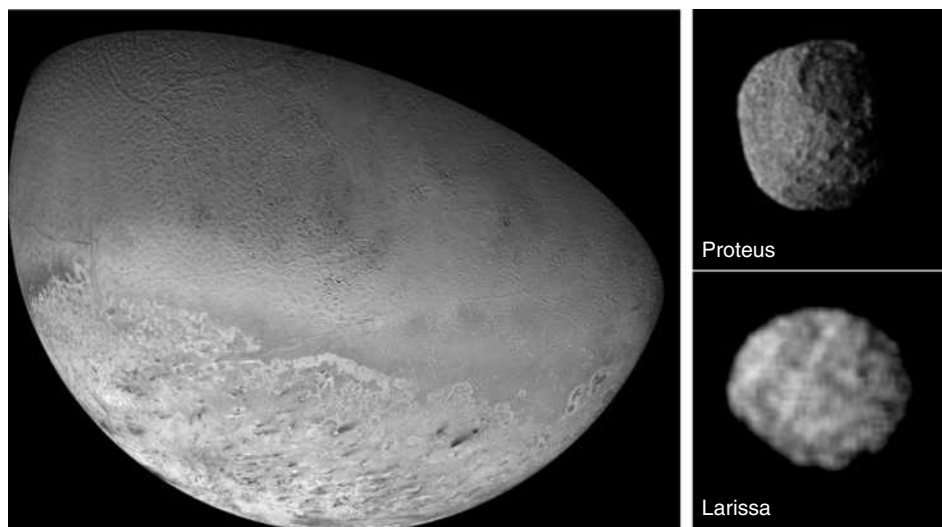


Figure 11. Voyager imaging of Triton and newly discovered satellites 1989N1 and 1989N2. (Left) [JPL P34687]. This Voyager 2 image has a resolution of 10 km. The South Pole, which is sunlit throughout the current season, is at bottom left. The absence of large impact craters suggests that Triton's surface has been renewed within the last billion years. (Upper right) [JPL P34727]. Voyager 2 image of Neptune's satellite 1989N1 (Proteus) at a resolution of 2.7 km. Its average diameter is 208 km. Its albedo is only 6%, compared to Triton's 76%, and its color is gray. (Lower right) [JPL P34698]. Voyager 2 image of 1989N2 (Larissa), Neptune's fourth largest satellite (mean radius 95 km), at a resolution of 4.2 km. It also has a low albedo (about 5%) and seems to have craters 30–50 km in diameter. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

that it may be a captured object rather than having formed in place. Nereid, though smaller than Proteus, is somewhat brighter. It was discovered by Kuiper in 1949. Nereid's photometric properties suggest that it has a surface of dirty frost. Of these comparably sized satellites, only Proteus has been imaged well enough to permit even a crude map of surface features. Its most prominent feature is an impact basin (72) about 210 km in diameter (larger than the 208-km radius of Proteus). The capture of Triton (see next section) and its evolution would have greatly perturbed any inner satellites, inducing mutual collisions and breakup. Catastrophic disruption would also be likely from external sources. Thus, the inner satellites are likely to be reaccreted debris. It remains unexplained why Nereid is brighter than the inner satellites. One possibility is that its greater distance prevented it from acquiring a veneer of dark particles lost from the rings. Among the known satellites, only Galatea seems to have any dynamic influence on Neptune's rings, which is to confine the ring arcs azimuthally (73), as discussed in a later section.

Triton. Neptune's satellite, Triton (Figs. 11 and 12), is one of the most peculiar of all satellites. It is the only satellite that has a retrograde orbit (it orbits opposite to the direction of Neptune's rotation). It was discovered by William Lassell, an amateur astronomer, less than a month after the discovery of Neptune — no

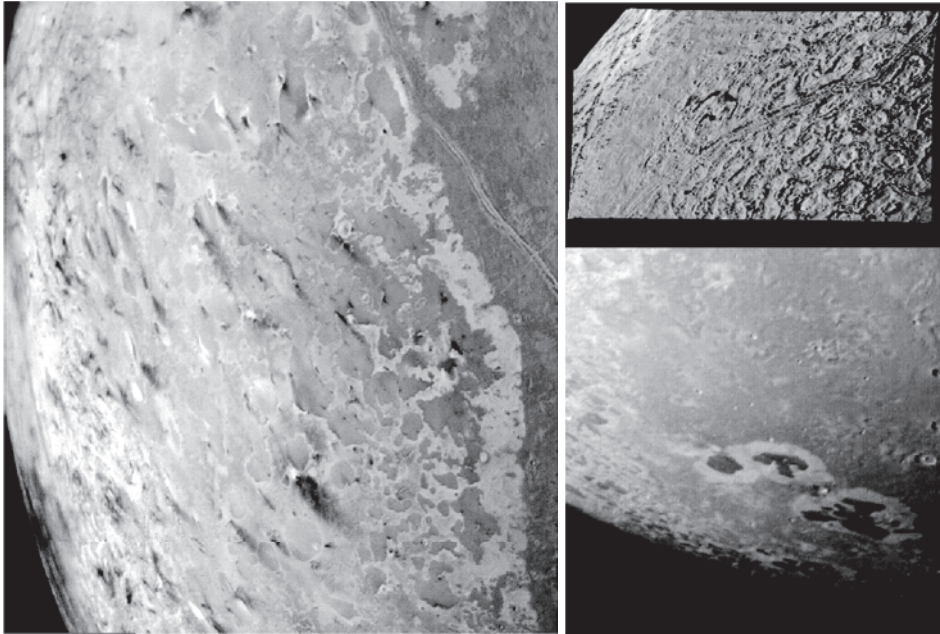


Figure 12. Close-up views of Triton. (Left) [JPL P34714]. Voyager 2 1989 image of the south polar terrain of Triton, showing 50 dark plume deposits or “wind streaks” on the icy surface. The plumes originate at dark spots a few miles in diameter, and some deposits stretch for more than 100 miles. A few active plumes were observed during the Voyager encounter. (Lower right) [JPL P34690]. Voyager 2 image of irregular dark patches on Triton’s surface. (Upper right) [JPL P34722]. Voyager 2 image of Triton’s cantaloupe-like terrain at a resolution of about 750 m. This terrain form of roughly circular depressions separated by rugged ridges is unique to Triton and covers large areas in its Northern Hemisphere.

mean feat given that it is 200 times fainter than Neptune! Prior to the Voyager encounter with Neptune, there was considerable speculation about the size of Triton and whether or not it had an atmosphere. From its deflection of Voyager’s orbit and imaging of its surface, the size and mass were finally established in 1989. Its diameter of 2710 km is larger than Pluto’s 2300 km, though smaller than Titan’s 3150 km. Its surface albedo of 72% at visible wavelengths (74) was that of an icy surface, but its density of 2.05 grams/cm^3 implied that rocky material also had to be a significant component. This density is comparable to that of the Pluto/Charon system. Only a very tenuous atmosphere of mainly nitrogen was observed and a surface pressure of only about 15 microbars.

Voyager observations revealed slightly pinkish bright regions and slightly bluish to gray dark regions, but the composition of the surface materials could be obtained only from ground-based spectroscopic observations. In 1978 observations of Triton, Cruikshank and Silvaggio (75) identified the spectral signature of methane gas. Later observations identified spectral features indicating that frozen methane (NH_4), solid molecular nitrogen (N_2), carbon monoxide (CO), carbon dioxide (CO_2), and water (H_2O) were present on the surface. Because of their extreme volatility, the existence of N_2 , CO , and CH_4 in solid form on Triton’s

surface means that Triton must be extremely cold. In fact, Triton's high reflectivity and great distance from the Sun make it one of the coldest places in the solar system; it has a surface temperature of only 38 K. The small amount of vapor sublimed into the atmosphere at these temperatures (mainly N_2) is balanced by condensation from the atmosphere. The equilibrium point at which condensation equals sublimation occurs at a pressure of only 15 microbars (15 millionths the pressure at the surface of Earth). Slightly warmer and slightly colder regions upset this local balance and produce atmospheric pressure variations, winds, and a redistribution of surface materials. As Triton's seasons lead to variations in the distribution of solar heating, the distribution of Triton's surface materials is likely to change over time, along with its mean albedo, its light curve (its brightness variation as a function of rotational angle), and perhaps its color. In fact, Triton's color in 1989 seems to have been considerably less red than it was in 1979 (51).

Voyager images of Triton's limb revealed discrete clouds and thin hazes at altitudes up to 30 km (51). The haze particles are possibly CH_4 ice and/or more complex organics created by interactions with UV radiation. Organic compounds may also be responsible for some of the slightly darker material on the surface, although rocky materials are also plausible contributors to surface features. Some of the most remarkable images of Triton were those of the active geyser-like plumes. The geysers were about 8 km tall, rising vertically from a small spot on a surface then abruptly ending in a dark cloud that extended in a narrow (about 5 km wide) plume downwind for as much as 150 km before disappearing (51). The cloud form suggests that horizontal wind speed increases abruptly at 8 km. The plumes are a plausible cause of the numerous dark streaks seen in the south polar region (Fig. 12). Venting of nitrogen gas, or perhaps methane, induced by solar heating (most active plumes are near subsolar latitudes) or localized geothermal energy, seems to entrain small dark particles that form the plume. The solar driven model involves dark material deposited under a layer of transparent frozen nitrogen that becomes heated by the Sun to a point at which the nitrogen in contact with the dark material is vaporized, builds in pressure, and eventually is released through a crack or rupture in the surface.

Triton's seasons are extreme as a result of its inclined and precessing retrograde orbit. Neptune's spin axis is tipped 29° relative to its orbital plane normal (compared to Earth's 23.5° tilt relative to its orbit normal). If Triton orbited within Neptune's equatorial plane, it would have seasonal forcing similar to Neptune's (and similar to Earth's as well, though much reduced in amplitude because of its great distance from the Sun). Triton's orbit, however, is inclined 21° from Neptune's equatorial plane, and because of the torque exerted by Neptune's oblate mass distribution, it precesses in a period of 688 years (76). That inclination can either add to or subtract from Neptune's. At one extreme, Triton's rotational (and orbital) axis can be tipped by $29^\circ + 21^\circ = 50^\circ$ away from the Neptune orbit normal, putting much of one of Triton's hemisphere in constant darkness throughout its 5.9-day orbital period and leading to heating at latitudes constantly exposed to the Sun and cooling at latitudes that are in constant darkness. This results in a transfer of volatile materials from the heated region to the cooled region, forming a polar cap of fresh ice at one pole and eroding the cap that had previously formed at the other pole. At another extreme, Triton's

rotational axis can precess to a point at which it has only an 8° ($29^\circ - 21^\circ$) inclination to the Neptune orbit normal and thus would receive a much more uniform exposure to solar heating in Northern and Southern Hemispheres. As Neptune orbits the Sun every 164.8 years, the subsolar latitude on Triton will be further modulated between positive and negative values of the current orbital inclination angle, leading to a two-component modulation of seasons on Triton; the shortest period of modulation is the 164.8 year orbital period of Neptune, which has an amplitude envelope modulated by the 688 year period of precession of Triton's orbital plane about Neptune. The summer solstice of 2000 will be followed by an equinox in 2041, during which Triton's surface should undergo considerable change.

Triton's surface is unlike that of any other satellite. Much of it appears roughened like the surface of a cantaloupe (upper right image of Fig. 12), formed by multitudes of circular dimples called cavi, which are about 25–30 km in diameter. The largest impact crater is only 27 km in diameter. From the small number of craters observed, Triton's surface appears to be relatively young geologically. It has large plains, apparently flooded by cryovolcanic fluids. Linear ridges 12–15 km wide indicate cracking of the surface and upwelling of material within the cracks (Figs. 11 and 12). Peculiar irregular blotches that have bright aureoles (Fig. 12, lower right) are of unknown origin.

Ring Systems

Uranus has a much more extensive and massive ring system than Neptune, perhaps a consequence of the additional debris generated by the larger impact event that presumably tilted Uranus' spin axis.

Rings of Uranus. The basic characteristics of the rings of Uranus are summarized in Table 5. Rings were first discovered by stellar occultation (dimming of a star's brightness as rings pass between a star and an observer). By measuring the star's brightness as a function of time, it is possible to determine ring

Table 5. **Rings of Uranus^a**

Ring	Distance, km	Width, km	Optical depth	Albedo
1986U2R	38,000	2,500	<0.001	0.03
6	41,840	1–3	0.2–0.3	0.03
5	42,230	2–3	0.5–0.6	0.03
4	42,580	2–3	0.3	0.03
Alpha	44,720	7–12	0.3–0.4	0.03
Beta	45,670	7–12	0.2	0.03
Eta	47,190	0–2	0.1–0.4	0.03
Gamma	47,630	1–4	1.3–2.3	0.03
Delta	48,290	3–9	0.3–0.4	0.03
1986U1R	50,020	1–2	0.1	0.03
Epsilon	51,140	20–100	0.5–2.1	0.03

^aRef. 69. Distance is from Uranus' center to the ring's inner edge.

locations and optical depth. In 1977, nine rings were discovered and characterized using stellar occultation (77). These are unofficially named 6, 5, 4, alpha, beta, eta, gamma, delta, and epsilon, in order of increasing distance from Uranus. Samples of the 1986 Voyager imaging of the Uranian rings are displayed in Fig. 13. The left image compares the image in reflected light on approach to Neptune as the image is in forward scattered light, taken after Voyager passed Uranus. Rings that have a significant component of small particles (of the order of a micron) appear much brighter in forward scattering (much like dust on a car windshield). Large particles dominate the appearance in backscattered light. (See the figure caption for further details of this comparison.)

The Uranian rings are narrow and sharp-edged and have optical depths of 0.3 or more. Most are very narrow (no more than 10 km wide), inclined to the equatorial plane of Uranus, and eccentric. Exceptions include the gamma and epsilon rings, which are not inclined, and the eta ring, which is not inclined and nearly circular. The epsilon ring is the widest and brightest (the most prominent ring in Fig. 5) and is also the most eccentric. It varies in distance from Uranus by about 800 km, and varies in width from 20 km where it is closest to Uranus, to 100 km where it is furthest. Its variation in radial distance is five times that of the next most eccentric ring (ring 5). It is somewhat of a mystery why orbital speed differences across the epsilon ring do not spread the ring material radially, though it is likely that some satellite resonances are responsible. Two shepherd satellites have been identified for the epsilon ring (right image of Fig. 13). These provide forces tending to confine the ring particles radially. The alpha and beta rings also vary systematically in width, from 5 km to 12 km; extremes are offset

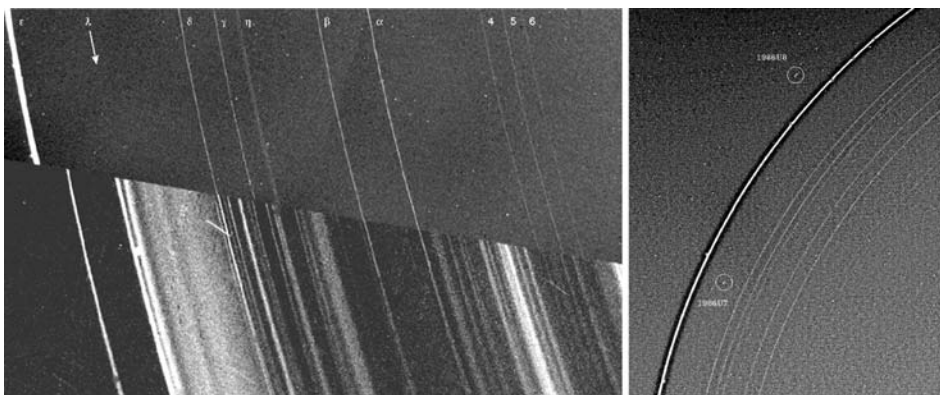


Figure 13. (Left) (Fig. 16–14 of Ref. 78). Voyager 2 images of the Uranian ring system first in backscattered light (upper half) one day before passing by Uranus in January 1986 and second, in forward scattered light (lower half), taken after passing by Uranus and looking backward. The nine labeled rings (upper half) are those discovered by stellar occultation measurements. The forward scattering view dramatically enhances the visibility of micron-sized particles, revealing structures not otherwise visible. Also note the mismatch of the two views of the epsilon ring, a consequence of its significant eccentricity. (Right) [JPL P29466]. Voyager 2 discovered two moons (1986U7, named Cordelia, and 1986U8, named Ophelia) that are shepherd satellites. The inner moon pushes ring particles outward, and the outer moon pushes them inward. This prevents the narrow rings from spreading out.

by about 30° in orbital longitude from the closest and most distant positions (periapsis and apoapsis). The most inclined rings (6, 5, and 4) deviate from Uranus' equatorial plane by only 24 to 46 km. A much more complex series of rings is seen in forward scattered light. In Fig. 13, there is an obvious lack of correlation between regions of high dust density (bright in forward scattering due to wavelength-sized particles) and regions of large particles (seen in backscatter and occultation observations). The structure and lack of correlation between dust features and large-particle features are reminiscent of Saturn's D ring.

Rings of Neptune. Neptune's rings (Table 6 and Fig. 14) are the least understood. Following the discovery of rings around Uranus, stellar occultations were also used to search for rings around Neptune. After several failures, the first detection in 1984 puzzled astronomers by showing rings on only one side of the planet. Voyager imaging in 1989 revealed that though Neptune's rings did completely encircle the planet, Neptune's ring particles were not uniformly distributed along the rings. Instead, much of the ring mass was clumped in restricted ring arcs. The outermost (Adams) ring, though continuous, contained three main arcs of much higher particle density of the order of 10° wide in longitude. The confinement of material in the ring arcs, is thought to be the result of gravitational interactions with the moon Galatea (80). One resonance interaction confines the material radially, producing narrow rings, and a second resonance interaction produces clumping of the ring material longitudinally, although the theory predicts regular spacing of clumps. It is not clear why the material is not periodically clumped. Only two rings are prominent in images taken on the sunlit side.

The ring material is very dark and probably red (80). The Neptune ring particles are as dark as those in the rings of Uranus, and have a single scattering albedo of about 0.04. A plausible composition is ice mixed with silicates and/or some carbon-bearing material. The Adams and Le Verrier rings contain a significant fraction of dust, comparable to the fraction in Saturn's F ring or Jupiter's ring; both are significantly dustier than the main rings of Saturn or Uranus.

Table 6. **Rings of Neptune**

Ring	Distance (km)	Width (km)	Optical depth ^b	Albedo ^b
Galle	41,900	15 ^a ~ 2000 ^b	~ 0.00008	~ 0.015
Le Verrier	53,200	15 ^a ~ 110 ^b	0.01–0.02 ^a ~ 0.002 ^b	~ 0.015
Lassell	53,200–57,200	5800 ^a ~ 4000 ^b	0.0001 ^a ~ 0.00015 ^b	~ 0.015
Arago ^b	57,200	< ~ 100 ^b		
Unnamed ^b	61,950			
1989N1R Adams	62,930	< 50 ^a ~ 50 ^b	0.01–0.1 ^a 0.0045 ^b 0.12(arcs) ^b	~ 0.015 0.04 (arcs) Egalité 1, Egalité 2)

^aRef. 69. Distance is from Uranus' center to the ring's inner edge.

^bRef. 79.

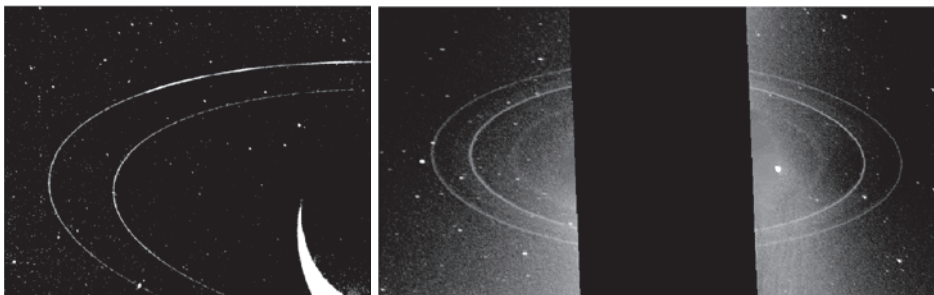


Figure 14. (Left) [JPL P35060]. Three clumps, or ring arcs, are visible in this view of Neptune's outermost Adams ring, imaged by Voyager 2 in August 1989. (Right) [JPL P35023]. This pair of Voyager 2 clear-filter images shows the ring system at the highest sensitivity and at a phase angle of 134° . The brighter, narrow rings are the Adams and Le Verrier rings. Extending out from the Le Verrier ring is the diffuse Lassell ring. The inner medium width ring is the Galle ring. The ring arcs seen in the left-hand image were in the blacked out region between these more sensitive images.

Satellites disrupted by collisions provide a plausible model for the source of Neptune's ring material. Meteorite collisions seem to be sufficient to explain the dust in the main Uranian rings (81), and in the diffuse Galle and Lassell rings (impacting some unseen parent body), but a more prolific source (perhaps interparticle collisions) is needed to account for the high abundance of dust in the Le Verrier and Adams rings.

Magnetic Fields and Magnetospheres

Jupiter's magnetic field was obvious from the ground because of the synchrotron radiation that it generated, but the magnetic fields of Uranus and Neptune were substantiated only by Voyager observations during close approaches to the planets. The supersonic solar wind particles interact with planetary fields to form a bow shock, and inside the bow shock, the solar wind encounters and is deflected by the planetary field, forming a boundary called the magnetopause. For Uranus, the detached bow shock was observed as a sudden increase in magnetic field intensity upstream at 23.7 Uranus radii (or $23.7 R_U$), and the magnetopause boundary was seen at $18 R_U$ (82). For Neptune, the corresponding distances (in Neptune radii) were $34.9 R_N$ and $26.5 R_N$. The magnetospheres are not complete barriers to solar wind particles, however, especially in the polar regions where particles are able to impact the atmosphere and generate an aurora. The UV spectrometer of Voyager 2 found auroral emissions from both Uranus and Neptune. The large offsets and tilts of the magnetic fields (discussed later) lead to auroral zones far from the poles of the rotational axes. The solar wind particles also contribute charged particles that form radiation belts. The Uranian rings and moons are embedded deep within the magnetosphere and thus play a role as absorbers of trapped radiation belt particles, as confirmed for Uranus by depressions in electron counts at magnetic latitudes swept by Miranda, Ariel, and Umbriel (83). A similar situation exists for Neptune's rings and satellites, except

for Nereid, which is outside Neptune's magnetosphere when it is on the sunward side of the planet.

The magnetic fields of Uranus and Neptune have unusually large inclinations relative to their rotational poles: the north dipole inclination is 58.6° for Uranus (84) and 47° for Neptune (85). The dipole moments are 48 and 25 times that of Earth (which is 7.9×10^{25} gauss cm^3). For comparison, Jupiter's moment is 20,000 times that of Earth, Saturn's 600 times, and respective inclinations are 9.6° and less than 1° , compared to Earth's 11° . Besides inclination relative to the spin axis, these magnetic dipoles are also unusual in having large offsets from the planet centers (Fig. 15) by $0.3 R_U$ for Uranus and $0.55 R_N$ for Neptune. However, the magnetic field of Uranus is not purely that of a dipole. It also has a strong quadrupole component, which is most significant close to the planet and contributes to the large variability of the magnetic field near the cloud tops that ranges from 0.1 gauss on the 1986 sunlit hemisphere to 1 gauss at a point on the 1986 dark hemisphere (Earth's field is about 0.3 gauss near the equator). The rotational periods of the magnetic fields are used to define the rotational periods of the interiors, which are given in Table 1.

The primary mechanism for producing magnetic fields in major planets, it is thought, is the dynamo mechanism, which appears to have three basic requirements (86): (1) planetary rotation, (2) a fluid electrically conducting region of the interior, and (3) convection within the conducting fluid. For Jupiter and Saturn, the rotation is rapid, the conducting fluid is metallic hydrogen, and primordial

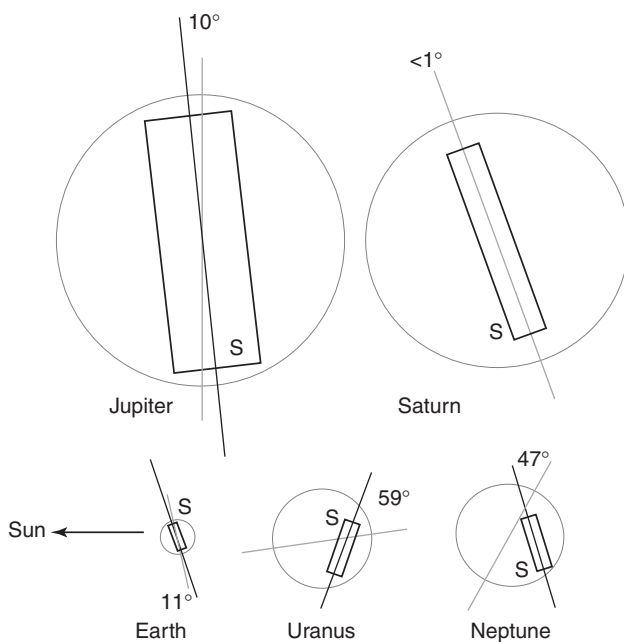


Figure 15. Orientations and offsets of outer planet magnetic fields. The north pole of the rotational axis is at the upper end of each indicated rotational axis, and the orbital planes are horizontal. The magnetic dipole locations are indicated by rectangles; the magnetic south end is indicated by S.

heat of formation drives the interior convection. For Uranus and Neptune, the rotations are somewhat slower, and the conducting fluid appears to be hot “ices” composed of water, methane, ammonia, and compounds derived from them. Neptune has a measurable internal heat flux exceeding the solar input that can drive interior convection, but Uranus has no measurable internal heat flux. Yet, Uranus actually has the larger dipole moment!

BIBLIOGRAPHY

1. Seidelmann, P.K. (ed.). *Explanatory Supplement to the Astronomical Almanac*. University Science Books, Mill Valley, CA, 1992, p. 316, Table 5.581.
2. Davies, M.E., V.K. Abalakin, A. Brahic, M. Bursa, B.H. Chovitz, J.H. Lieske, P.K. Seidelmann, A.T. Sinclair, and I.S. Tiuffin. Report of the IAU/IAG/COSPAR Working Group on cartographic coordinates and rotational elements of the planets and satellites—1991. *Celestial Mech. Dynamical Astron.* 53: 377–397 (1992)
3. Beatty, J.K., C.C. Petersen, and A. Chaikin (eds). *The New Solar System*, 4th ed. Sky Publishing and Cambridge University Press, Cambridge, UK, 1999, p. 387.
4. Forbes, E.G. Pre-discovery observations of Uranus. In *Uranus and the Outer Planets*, Proceedings of the IAU/RAS colloquium no. 60, Cambridge University Press, Cambridge, UK, 1982, pp. 67–79.
5. Moore, P. The discoveries of Neptune and Triton. In *Neptune and Triton*, D.P. Cruikshank (ed.). University of Arizona Press, Tucson, AZ, 1995, pp. 15–36.
6. Wood, J.A. Origin of the solar system. In *The New Solar System*, 4th ed. J. Beatty, C. Petersen, and A. Chaikin (eds). Cambridge University Press, Cambridge, England, 1999.
7. Hubbard, W.B. Interiors of the giant planets. In *The New Solar System*, 4th ed. J. Beatty, C. Petersen, and A. Chaikin (eds). Cambridge University Press, Cambridge, England, 1999.
8. Lissauer, J.J., J.B. Pollack, G.W. Wetherill, and D.J. Stevenson. Formation of the Neptune system. In *Neptune and Triton*, D.P. Cruikshank (ed.). University of Arizona Press, Tucson, AZ, 1995, pp. 37–108.
9. Lissauer, J.J., and V.S. Safronov. The random component of planetary rotation. *Icarus* 93: 288–297 (1991).
10. Hubbard, W.B., M. Podolak, and D.J. Stevenson. The interior of Neptune. In *Neptune and Triton*, D.P. Cruikshank (ed.). University of Arizona Press, Tucson, AZ, 1995, pp. 109–138.
11. Lewis, J.S. *Physics and Chemistry of the Solar System*. Academic Press, San Diego, 1997.
12. Pearl, J.C., and B.J. Conrath. The albedo, effective temperature, and energy balance of Neptune, as determined from Voyager data. *J. Geophys. Res.* 96: 18921–18930 (1991).
13. Holme, R., and A.P. Ingersoll. Baroclinic instability in the interiors of the giant planets: A cooling history of Uranus? *Icarus* 110(2): 340–356 (1994).
14. Ingersoll, A.J. Atmospheres of the giant planets. In *The New Solar System*, 4th ed. J. Beatty, C. Petersen, and A. Chaikin (eds). Cambridge University Press, Cambridge, England, 1999, pp. 201–220.
15. Kuiper, G.P. Planetary atmospheres and their origin. In *The Atmospheres of the Earth and Planets*, G. Kuiper (ed.). University of Chicago Press, Chicago, 1952, pp. 306–405.
16. Hertzberg, G. Spectroscopic evidence of molecular hydrogen in the atmospheres of Uranus and Neptune. *Astrophys. J.* 115: 337–340 (1952).

17. Bergstrahl, J.T., E.D. Miner, and M.S. Matthews (eds). *Uranus*. University of Arizona Press, Tucson, AZ, 1991.
18. Gierasch, P.G., and B.J. Conrath. Vertical temperature gradients on Uranus: Implications for layered convection. *J. Geophys. Res.* 92: 15019–15029 (1987).
19. Flasar, F.M. Para-hydrogen equilibrium and superadiabatic lapse rates on Uranus: Evidence of methane condensation? *Bull. Am. Astron. Soc.* 19: 757 (1986).
20. Conrath, B.R., Hanel, D. Gautier, A. Martin, and G. Lindal. The helium abundance of Uranus from Voyager measurements. *J. Geophys. Res.* 92: 15003–15010 (1987).
21. Conrath, B.J., D. Gautier, A. Martin, G.F. Lindal, R.E. Samuelson, and W.A. Shaffer. The helium abundance of Neptune from Voyager measurements. *J. Geophys. Res.* 96: 18907–18919 (1991).
22. Baines, K.H., H.B. Hammel, K.A. Rages, P.N. Romani, and R.E. Samuelson. Clouds and hazes in the atmosphere of Neptune. In *Neptune and Triton*, D.P. Cruikshank (ed.). University of Arizona Press, Tucson, AZ, 1995, pp. 489–546.
23. Yelle, R.V., F. Herbert, B.R. Sandel, R.J. Vervack, Jr., and T.M. Wentzel. The distribution of hydrocarbons in Neptune's upper atmosphere. *Icarus* 104: 38–59 (1994).
24. Fegley, B.F., D. Gautier, T. Owen, and R.G. Prinn. Spectroscopy and chemistry of the atmosphere of Uranus. In *Uranus*, J.T. Bergstrahl, E.D. Miner, and M.S. Matthews. (eds). University of Arizona Press, Tucson, AZ, 1991, pp. 147–203.
25. Lewis, J.J., and R.G. Prinn. Kinetic inhibition of CO and N₂ reduction in the solar nebula. *Astrophys. J.* 238: 357–364 (1980).
26. Karkoschka, E. Spectrophotometry of the jovian planets and Titan at 300- to 1000-nm wavelength: The methane spectrum. *Icarus* 111: 174–192 (1994).
27. Fink, U., and S. Larson. The infrared spectra of Uranus, Neptune, and Titan from 0.8 to 2.5 microns. *Astrophys. J.* 233: 1021–1040 (1979).
28. Sromovsky, L.A., P.M. Fry, S.S. Limaye, K.H. Baines, and T. Dowling. HST and IRTF Observations of Neptune during 1998. *Bull. Am. Astron. Soc.* 30: 1098 (1998).
29. Sromovsky, L.A., P.M. Fry, K.H. Baines, and T.E. Dowling. Coordinated imaging of Neptune and Triton: II. Implications of disk-integrated photometry. *Icarus* 149: 459–488 (2001).
30. Sromovsky, L.A., P.M. Fry, T.E. Dowling, K.H. Baines, and S.S. Limaye. Neptune's atmospheric circulation and cloud morphology: Changes revealed by 1998 HST imaging. *Icarus* 150: 244–260 (2001).
31. Karkoschka, E. Clouds of high contrast on Uranus. *Science* 280: 870–872 (1998).
32. Allison, M., and L. Travis (eds). *The Jovian Atmospheres*. NASA CP 2441, 1986.
33. Conrath, B.J., J.C. Pearl, J.F. Appleby, F.F. Lindal, G.S. Orton, and B. Bezard. Thermal structure and energy balance of Uranus. In *Uranus*, J.T. Bergstrahl, E.D. Miner, and M.S. Matthews. (eds). University of Arizona Press, Tucson, AZ, 1991, pp. 204–252.
34. Moses, J.I., M. Allen, and Y.L. Yung. Hydrocarbon nucleation and aerosol formation in Neptune's atmosphere. *Icarus* 99: 318–346 (1992).
35. Lindal, G.F., J.R. Lyons, D.N. Sweetnam, V.R. Eshelman, D.P. Hinson, and G.L. Tyler. The atmosphere of Uranus: Results of radio occultation measurements with Voyager 2. *J. Geophys. Res.* 92: 14987–15001 (1987).
36. Lindal, G.F. The atmosphere of Neptune: An analysis of radio occultation data acquired with Voyager 2. *Astron. J.* 103: 967–982 (1992).
37. Romani, P.N., I. de Pater, and S.K. Atreya. Neptune's deep atmosphere revealed. *Geophys. Res. Lett.* 16: 933–936 (1989).
38. Romani, P. N., J. Bishop, B. Bezard, and S. K. Atreya. Methane photochemistry on Neptune: Ethane and acetylene mixing ratios and haze production. *Icarus* 106: 442–463 (1993).
39. Baines, K.H., and H.B. Hammel. Clouds, hazes, and the stratospheric methane abundance in Neptune. *Icarus* 109: 20–39 (1994).

40. Baines, K.H., M.E. Mickelson, L.E. Larson, and D.W. Ferguson. The abundances of methane and ortho/para hydrogen in Uranus and Neptune: Implications of new laboratory 4-O H₂ quadrupole line parameters. *Icarus* 114: 328–340 (1995).
41. Walter, C.M., and M.S. Marley. The Uranian geometric albedo: An analysis of atmospheric scatterers in the near-infrared. *Icarus* 132: 285–297 (1998).
42. Rages, K., J.B. Pollack, M.G. Tomasko, and L.R. Dose. Properties of scatterers in the troposphere and lower stratosphere of Uranus based on Voyager imaging data. *Icarus* 89: 359–376 (1991).
43. West, R.A., K.H. Baines, and J.B. Pollack. Clouds and aerosols in the Uranian atmosphere. In J.T. Bergstrahl, E.D. Miner, and M.S. Matthews (eds). *Uranus*, University of Arizona Press, Tucson, AZ, 1991, pp. 296–324.
44. Gautier, D., B.J. Conrath, T. Owen, I. de Pater, and S.K. Atreya. The troposphere of Neptune. In D.P. Cruikshank (ed.). *Neptune and Triton*, University of Arizona Press, Tucson, AZ, 1995.
45. de Pater, I., and D.L. Mitchell. Microwave observations of the planets: The importance of laboratory measurements. *J. Geophys. Res.* 91: 220–233 (1993).
46. Crisp, D., J. Trauger, K. Stapelfeldt, T. Brooke, J. Clarke, G. Ballester, and R. Evans. Hubble Space Telescope Wide Field Planetary Camera 2 observations of Neptune. *Bull. Am. Astron. Soc.* 26: 1093–1093 (1994).
47. Sromovsky, L.A., S.S. Limaye, and P.M. Fry. Dynamics of Neptune’s major cloud features. *Icarus* 105: 110–141 (1993).
48. Sromovsky, L.A., P.M. Fry, and K.H. Baines. The unusual dynamics of northern dark spots on Neptune. *Icarus* 156: 16–36 (2002).
49. Polvani, L.M., J. Wisdom, E. DeJong, and A.P. Ingersoll. Simple dynamical models of Neptune’s Great Dark Spot. *Science* 249: 1393–1398 (1990).
50. LeBeau, R.P., and T.E. Dowling. EPIC simulations of time-dependent three-dimensional vortices with application to Neptune’s Great Dark Spots. *Icarus* 132: 239–265 (1998).
51. Smith, B.A., et al. Voyager 2 at Neptune: Imaging science results. *Science* 246: 1422–1449 (1989).
52. Stratman, P.W., A.P. Showman, T.E. Dowling, and L.A. Sromovsky. EPIC simulations of bright companions to Neptune’s Great Dark Spots. *Icarus* 151: 275–285 (2001).
53. Hammel, H.B., and G.W. Lockwood. Atmospheric structure of Neptune in 1994, 1995, and 1996: HST imaging at multiple wavelengths. *Icarus* 129: 466–481 (1997).
54. Sromovsky, L.A., P.M. Fry, T.E. Dowling, K.H. Baines, and S.S. Limaye. Coordinated imaging of Neptune and Triton: III. Neptune’s atmospheric circulation and cloud structure. *Icarus* 149: 459–488 (2001).
55. Hammel, H.B., K.A. Rages, and G.W. Lockwood. New observations of zonal winds on Uranus. *Bull. Am. Astron. Soc.* 30: 1097 (1998).
56. Sromovsky, L.A., J.R. Spencer, K.H. Baines, and P.M. Fry. Ground-based observations of cloud features on Uranus. *Icarus* 146: 307–311 (2000).
57. Lockwood, G.W., B.L. Lutz, D.T. Thompson, and A. Warnock III. The albedo of Uranus. *Astrophys. J.* 266: 402–414 (1983).
58. Karkoschka, E. Uranus’ apparent seasonal variability in 25 HST Filters. *Icarus* 151: 84–92 (2001).
59. Lockwood, G.W., and D.T. Thompson. Photometric variability of Neptune, 1972–2000. *Icarus* 156: 37–51 (2002).
60. Sromovsky, L.A., S.S. Limaye, and P.M. Fry. Clouds and circulation on Neptune: Implications of 1991 HST observations. *Icarus* 118: 25–38 (1995).
61. Hammel, H.B., S.L. Lawson, J. Harrington, G.W. Lockwood, D.T. Thompson, and C. Swift. An atmospheric outburst on Neptune from 1986 through 1989. *Icarus* 99: 363–367 (1992).

62. Joyce, R.R., C.L.B. Pilcher, D.P. Cruikshank, and D. Morrison. Evidence for Weather on Neptune I. *Astrophys. J.* 214: 657–662 (1977).
63. Hammel, H.B., K. Rages, G.W. Lockwood, E. Karkoschka, and I. de Pater, New measurements of the winds of Uranus. *Icarus* 153: 229–235 (2001).
64. Limaye, S.S., and L.A. Sromovsky. Winds of Neptune: Voyager observations of cloud motions. *J. Geophys. Res.* 96: 18941–18960 (1991).
65. Conrath, B.J., R.A. Hanel, and R.E. Samuelson. Thermal structure and heat balance of the outer planets. In *Origin and Evolution of Planetary and Satellite Atmospheres*, S.K. Atreya, J.B. Pollack, and M.S. Mathews (eds). Univ. of Arizona Press, Tucson, AZ, 1989, pp. 513–538.
66. Ingersoll, A.J., C.D. Barnet, R.F. Beebe, F.M. Flasar, D.P. Hinson, S.S. Limaye, L.A. Sromovsky, and V.E. Suomi. Dynamic meteorology of Neptune. In *Neptune and Triton*, D.P. Cruikshank (ed.). University of Arizona Press, Tucson, AZ, 1995, pp. 613–682.
67. Friedson, J., and A.P. Ingersoll. Seasonal meridional energy balance and thermal structure of the atmosphere of Uranus: A radiative-convective-dynamical model. *Icarus* 69: 135–156 (1987).
68. Compilation by JPL's Solar System Dynamics group, 24 April 2002, available at <http://ssd.jpl.nasa.gov>.
69. Beatty, J.K., et al. (eds). *The New Solar System*, 3rd ed. Sky Publishing and Cambridge University Press, Cambridge, UK, 1990.
70. Smith, B.A., et al. Voyager 2 in the Uranian system: Imaging science results. *Science* 233: 43–64 (1986).
71. Thomas, P.C., J. Veverka, and P. Helfenstein. Voyager observations of Nereid. *J. Geophys. Res.* 96: 19261–19268 (1991).
72. Thomas, P.C., J. Veverka, and P. Helfenstein. Neptune's small satellites. In *Neptune and Triton*, D.P. Cruikshank, (ed.). University of Arizona Press, Tucson, AZ, 1995.
73. Porco, C. An explanation for Neptune's ring arcs. *Science* 235: 995–1001 (1991).
74. Cruikshank, D.P. Triton, Pluto, and Charon. In *The New Solar System*, J.K. Beatty, C.C. Petersen, and A. Chaikin, (eds). Sky Publishing and Cambridge University Press, Cambridge, UK, 1999, pp. 285–310.
75. Cruikshank, D.P., and P.M. Silvaggio, Triton: A satellite with an atmosphere. *Astrophys. J.* 233: 1016–1020 (1979).
76. Harris, A.W. Physical properties of Neptune and Triton inferred from the orbit of Triton. In *Uranus and Neptune*, J. Bergstrahl (ed.). NASA CP 2330, 1984, pp. 357–373.
77. Elliot, J.L., and P.S. Nicholson. In *Planetary Rings*, R. Greenberg and A. Brahic (eds). University of Arizona Press, Tucson, AZ, 1984.
78. Burns, J. Planetary rings. In *The New Solar System*, J.K. Beatty, C.C. Petersen, A. Chaikin (eds). Sky Publishing and Cambridge University Press, Cambridge, UK, 1999, pp. 221–240.
79. Williams, D.R. Neptune Rings Fact Sheet, GSFC, National Space Science Data Center, available at <http://nssdc.gsfc.nasa.gov/planetary/factsheet/nepringfact.html>.
80. Porco, C.C., P.D. Nicholson, J.N. Cuzzi, J.J. Lissauer, and L.W. Esposito. Neptune's ring system. In *Neptune and Triton*, D.P. Cruikshank (ed.). University of Arizona Press, Tucson, AZ, 1995.
81. Colwell, J.E., and L.W. Esposito. A numerical model of the Uranian dust rings. *Icarus* 86: 530–560 (1990).
82. Ness, N.F., M.H. Acuna, K.W. Behannon, L.F. Burlaga, J.E.P. Connerny, R.P. Lepping, and F.M. Neubauer. Magnetic fields at Uranus. *Science* 233: 85–89 (1986).
83. Stone, E.C., J.F. Cooper, A.C. Cummings, F.B. McDonald, J.H. Trainor, N. Lal, R. McGuire, and D.L. Chenette. Energetic charged particles in the Uranian magnetosphere. *Science* 233: 93–97 (1986).

84. Ness, N.F., J.E.P. Connerney, R.P. Lepping, M. Schulz, and G.-H. Voigt, The magnetic field and magnetospheric configuration of Uranus. In *Uranus*, J.T. Bergstrahl, E.D. Miner, and M.S. Matthews (eds). University of Arizona Press, Tucson, AZ, 1991, pp. 739–779.
85. Ness, N.F., M.H. Acuna, L.F. Burlaga, J.E.P. Connerney, and R.P. Lepping. Magnetic fields at Neptune. *Science* 246: 1473–1478 (1989).
86. Van Allen, J.A., and F. Bagenal. Planetary magnetospheres and the interplanetary medium. In *The New Solar System*, J.K. Beatty, C.C. Petersen, A. Chaikin (eds). Sky Publishing and Cambridge University Press, Cambridge, UK, 1999, pp. 39–58.

OTHER READING

- Voyager 2 at Uranus*, Journal of Geophysical Research Special Issue, Vol. 92, no. A13, A.G.U., Washington, D.C., 1987.
- Miner, E. *Uranus: The Planet, Rings, and Satellites*, 2nd ed., Ellis Harwood, Chichester, UK, 1998.
- Hunt, G., and P. Moore. *Atlas of Uranus*. Cambridge University Press, Cambridge, England, 1989.
- Voyager 2 at Neptune*, Journal of Geophysical Research Supplement, Oct. 30, Vol. 96, A.G.U., Washington, D.C., 1991.
- Miner, E., and R. Wessen. *Neptune – The Planet, Rings, and Satellites*. Praxis, Chichester, England, 2002.
- Hunt, G., and P. Moore. *Atlas of Neptune*. Cambridge University Press, Cambridge, England, 1994.
- Moore, P. *The Planet Neptune*. Halsted Press, New York, 1988.
- Hunt, G.E. (ed.) *Uranus and the Outer Planets*. Proceedings of the IAU/RAS colloquium no. 60, Cambridge University Press, Cambridge, 1982.

LAWRENCE A. SROMOVSKY
University of Wisconsin
Madison, Wisconsin

V

VEGA PROJECT

Vega was the last successful deep space mission of the Soviet space program. It was built on the legacy of the Venera-9 to Venera-16 series of spacecraft launched, starting in 1975, to study Venus. In its last encounter (1986), Halley's comet's perihelion was much closer to Venus than to Earth. That created a unique opportunity to launch Vega during the astronomical window for Venus (December 1984) and to use that planet's gravity assist for redirecting the craft to the comet. Using such a choice of spacecraft trajectory, it was decided to combine a scenario to explore Venus with a Halley's comet encounter by employing a two-element space vehicle: a planetary reentry module, carrying the Venus Lander and balloon, and a Halley flyby probe. The integrated mission was called Vega, a contraction of the Russian words "Venera" (Venus) and "Gallei" (Halley) and was conducted by the Soviet Union and a number of other countries within the framework of an Interkosmos program.

The Vega mission comprised two identical spacecraft, Vega 1 and Vega 2. This was a standard approach in the Soviet Union, aimed primarily at increasing the overall reliability of the mission. In addition, if both flybys were successful, there would be a significant increase in the scientific return, which was particularly valuable considering the expected variability of the comet's activity.

The Vega project was truly international. The spacecraft production and launch operation were controlled by the Soviet aerospace agency (at that time called the Ministry of General Machine Building), but the scientific program and its payload were coordinated by the International Science and Technical Committee (ISTC), representing scientific institutions and space agencies from nine countries. The ISTC designed the Vega mission program to optimize international efforts related to the Halley's comet campaign, which included the European Giotto and the Japanese Suisei and Sakigake space missions.

The two spacecraft were launched by Proton rockets from the Baikonur launching site on 15 and 21 December 1984, respectively. On 11 and 15 June 1985 at the Venus flybys, the planetary packages were ejected and reentered the atmosphere of the planet (the scientific results are summarized in the Venus article elsewhere in this *Encyclopedia*). The delivery and deployment of balloons into the Venusian atmosphere was the first operation of this kind outside of the terrestrial atmosphere (and it stands alone to this day).

After Venus gravity assist maneuvers, the Vega 1 and Vega 2 trajectories were changed to encounter comet Halley on 6 and 9 March 1986, respectively.

Each Vega spacecraft, comprised of a Halley flyby probe and a Venus descent module, weighed about 4.5 t at launch. The spacecraft configuration at the beginning of its interplanetary journey is shown in the center of Fig. 1. The Venus descent module, ejected at the planet's flyby, had a package carrying the balloon, a minigondola (that had a radio transmitter, a basic atmospheric science package, and a lithium battery power source), and a pressurized helium tank which was mounted on the upper part of the descent module. The inflated balloon is presented on the right side of Fig. 1. The spacecraft was triaxially stabilized and had a span of solar panels of about 10 m. It carried 14 experiments, subdivided into three topical groups:

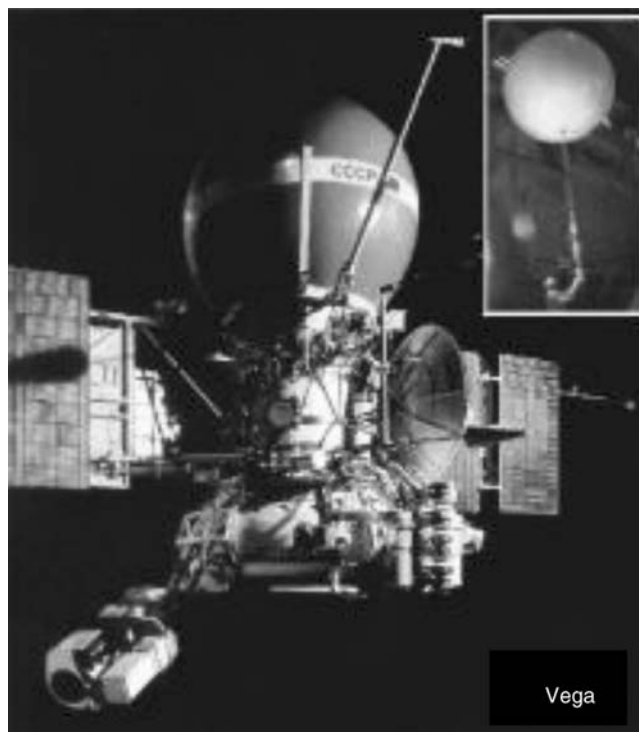


Figure 1. Vega configuration. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

- a space physics package for studying the solar wind versus cometary plasma interaction;
- a set of dust particle detectors for *in situ* measurements at the encounter with the comet;
- an imaging and spectral package of instruments, which included a TV system (TVS) for tracking and imaging the inner coma and the comet nucleus.

Taking the comet's coma images during the fast flyby (the relative velocity was close to 80 km/s) required mounting a steerable platform on the spacecraft body. This platform, shown in the lower part of Fig. 1, was mounted on the lower left side of the spacecraft, had a mass of 82 kg, and carried a 64-kg payload.

The Vega spacecraft design used as its baseline the Venera series of spacecraft. However, a number of modifications had to be introduced to improve the survivability of the probe in the harsh environment of the comet's coma. A special shield with a net surface area of about 5 m² was added to protect the most essential subsystems against the bombardment of dust particles. The design of such a shield underwent a series of tests, simulating the impact of projectiles at a velocity of 80 km/s by using a high-power pulsed laser. The results have confirmed computer simulation predictions for super high velocity impacts. Finally, a double-sheet bumper was adopted, composed of a thin metallic front sheet (0.4 mm) and a thicker rear sheet, separated by several centimeters.

The solar array had a total area of 10 m². Triaxial stabilization was achieved by a gyroscopic system in combination with a number of gas nozzles. The telemetry system consisted of a high-data-rate channel of 65,536 bps (BRL) and a low-data-rate channel of 3072 bps (BTM). The BRL channel was used only for real-time data transmission, including the data from the imaging instruments and the science payload. The housekeeping data were received via the BTM. The data could also be stored on board on a tape recorder and subsequently delivered by the BTM channel. During high-data-rate transmission periods, the spacecraft was oriented so that the high-gain antenna pointed toward Earth.

The imaging experiment on board the Vega spacecraft had the following scientific objectives:

- to detect and take images of the comet's nucleus (this was never done before, and Vega had to become the first to do it a few days before Giotto) and to determine its shape, size, and the albedo of surface material of the nucleus;
- to assess the composition and surface morphology of the nucleus;
- to identify the spatial pattern and temporal variation of its activity; and
- to study the structure and dynamics of the near-nucleus coma, including the jets of cometary material escaping from the nucleus.

The design of the imaging camera system for the Vega mission was based on using a Ritchey–Chrétien telescope that had a focal length of 1200 mm, an effective aperture of f 6.5 and a field of view of 26.4×39.6 minutes of arc. It was immediately obvious that the field of view of this narrow angle camera (NAC) was too small to locate the nucleus autonomously. It was also clear that ground

control could not provide the pointing commands for the telescope in real time. That task was assigned to the wide angle scanning camera (WAC), that served as an addition to the imaging payload.

The type of the imaging sensors also required careful consideration. Charge-coupled devices (CCDs) were just becoming available in the late 1970s for spaceflight application and only limited information could be obtained on their reliability in space. The Vega and Giotto missions were among the first space science missions that used this new technology. We were able to use CCDs for Vega that were manufactured in the Soviet Union.

The selection of the filters was also discussed extensively. Ideally, we would have liked to use narrowband filters for spectroscopy. However, the sensitivity of the CCDs available for the project was not high enough, and we had to select wideband filters. Due to the lack of advance knowledge of brightness conditions around the nucleus, the exposure time of the system was set up to provide under- and overexposed images in sequence. That assumed the image processing procedure described subsequently.

A major concern was identification of the comet nucleus, the main target for pointing the axis of the imaging instruments. The traditional “quadrant” type sensors, widely used in pointing technology, would integrate the light over an extended area around the target to find the center of brightness. Though quite appropriate for a point object, these sensors could be misled by the jets escaping from the comet’s nucleus. Being diffuse-type objects, they are not necessarily brighter than a nucleus, but considering their large extensions, the net signal could prevail. In sum, the center of brightness does not necessarily coincide with the position of the nucleus. The pointing system might well have locked on to a bright dust jet and been steered away from the nucleus. As there was no way to define an offset in advance and with sufficient reliability, the decision was to employ a special microprocessor, capable of performing a quick cluster analysis of images taken by the WAC in real flight to identify the brightest spot, attributed to the nucleus per se. In the end, this targeting strategy worked well for Vega.

The pointing platform was based on a precise servomechanism with two degrees of freedom. Final precision of pointing, achieved with such a platform even at its maximal angular velocity, ensured that smearing of images due to the motion of the platform and residual wobbling of the spacecraft body attitude was always less than 1 pixel. The pointing platform used on the spacecraft was designed and manufactured as a contribution from Czechoslovakia. Due to the absolutely critical importance of steering the platform for the success of the encounter with the comet, a major Soviet aerospace enterprise was assigned to build a backup version. Extensive tests of the Czech designed platform were sufficiently persuasive ultimately for deciding to use it on both spacecraft.

The default navigation system for platform pointing was the imaging system (WAC) itself, but Vega had a complementary means for backup. The second choice was an eight-segmented, light-sensitive sensor mounted on the pointing platform. Its functional principle was extremely simple: if the center of brightness of the comet moved away from the center of the sensor, a change in platform orientation would be initiated to compensate for this offset. The inner envelope of four sensors had to pick up at a closer flyby to reduce the offset. Fairly sophisticated computer simulations proved that this backup was reliable. This

algorithm performed well in an emergency when Vega 2 was close to the comet and the microprocessor in charge of pointing control was temporarily knocked out presumably by cosmic radiation.

Planning and manufacturing the telescope was also a major task. It was decided that this effort should be duplicated, so consequently, French and Soviet versions were prepared. Both worked satisfactorily, the French telescope was selected for Vega 1, and the Soviet telescope flew on board Vega 2.

To have critical backups was a major element in our design approach. A third navigational backup, which used analog signals from a completely independent CCD sensor, was added to the TV system. The output signal was processed similarly to the eight-segmented, light-sensitive sensor, but at much higher precision.

There was one major setback during camera development. We planned to use 8-bit analog-to-digital converters (ADCs) in the camera electronics. Suitable 8-bit ADCs were not available in Eastern Europe, and we had intended to buy them in the West. However, as we discovered only during the manufacturing phase, we were unable to buy space qualified 8-bit ADCs in the West due to technology transfer regulations. It was then decided to use the simpler Soviet design, which in the end could not provide equivalent contrast resolution. Unfortunately, this affected the scientific performance of the camera by damaging its contrast resolution.

After the encounters, both Vega spacecraft still remained largely functional despite exposure to intense bombardment by dust particles flying at a speed of 80 km/s. However, this did inflict some damage on the payload. The spacecraft attitude during the encounters was determined by the requirements to point the high-gain antenna at Earth and, at the same time, to obtain maximum power from the solar panels. Consequently, the solar panels could not be aligned with the relative-velocity vector, and damage to the panels was inevitable. After the encounters, the power from the solar panels was reduced by about 50%.

The deep space communication center was located in Evpatoria (Crimea, now part of Ukraine), and most of the science investigators worked on-line from Moscow at the Space Research Institute, where they were able to obtain all of the data in real time. Deep-space antennas in Evpatoria (70 m) and Medvezhy Ozera (64 m), near Moscow, received the telemetry data.

During the cruise phase, the pointing platform was clamped down. The clamping mechanism on Vega 1 was released on 14 February 1986 and that on Vega 2 on 18 February 1986, and the operation of the TVS and the pointing platform and imaging system were tested and calibrated by observing Jupiter and Saturn. Two days before the encounters, the cameras were oriented toward Halley's comet and switched on for 2 hours. At last, a few minutes before 7:20 UT on 6 March 1986, the nucleus of comet Halley was unveiled to the human eye for the first time in the history of humankind.

The second encounter took place almost exactly 3 days later. Both encounters were watched live by large audiences worldwide. Such openness was unprecedented for the Soviet space program.

The Vega mission operation was discontinued a few weeks after the encounters. The general condition of the spacecraft would have allowed further operation. The solar panels were partially damaged by dust impacts, but could

still have provided enough power. The camera performance was tested by observing Jupiter, and no essential degradation was registered. Even the amount of propellant on board was considerable. However, after extensive searches and debates, no interesting object was identified for a possible second encounter. Image processing commenced immediately after the flybys and was carried out in parallel by four cooperating teams located in Moscow, Budapest, Paris, and Berlin.

The Vega mission to Halley's comet was scientifically highly successful, and it also helped to forge international cooperation in space science at an unprecedented level, both within the Vega Project and in terms of worldwide cooperation with the other major space organizations. Together with the Giotto project, it led to the formation of the Inter-Agency Consultative Group for Space Science (IACG), which coordinated all major Halley's comet efforts by several Space Agencies: Intercosmos, NASA, ESA and ISAS (Japan) from 1981 to 1986. Following that initiative, the IACG continued international coordination for some other projects, even after 1986.

The best example of the usefulness of this cooperation was the "Pathfinder Concept." The idea was to target ESA's Giotto spacecraft as precisely as possible using of the Halley nucleus observations from the two earlier arriving Vega spacecraft. Before the encounters, the Halley nucleus position was known only with an accuracy of about a few hundred kilometers. That was all that could be achieved by combining ground-based astronomical observations over a long period with a model, which included nongravitational forces on the nucleus. This was sufficient for the Vega spacecraft, flying by some 10,000 km away, but not for Giotto. For Giotto, which was supposed to pass the nucleus at a distance of only 500 km on the sunward side, significantly better targeting uncertainty was highly desirable. Fortunately, Giotto had to be the last of the spacecraft to encounter the comet, and the information obtained by the cameras on board the earlier arriving Vega 1 and Vega 2 spacecraft could be passed on from the Vega to the Giotto Project. This was to become known as the "Pathfinder Concept."

Based on Pathfinder data, Giotto had to make its last fine-tuned orbit correction at least 2 days before the encounter on 11 March. Vega 2 encountered Halley on 9 March, so there was very little time between the evaluation of the Vega Pathfinder data, which included positional determination of both Vega spacecraft; cross-checking by three different groups at Moscow, Pasadena, and Darmstadt; and executing Giotto's maneuver. Both the Giotto and the Vega spacecraft uncertainties contributed to the error. Using conventional (6 GHz) ranging and Doppler techniques, the technical means available to Soviets expected a geocentric Vega positional accuracy of several hundred kilometers. Using precise L-band very long baseline interferometry (VLBI) techniques and NASA's widely separated tracking stations of the Deep Space Network (DSN), the positional uncertainties of both spacecraft could be reduced to about 40 km. After processing all of the data, Giotto was finally aimed at 540 km on the sunward side and actually achieved a flyby distance of 596 km. As a result, the Giotto camera was able to get a truly high resolution image of the Halley's comet nucleus.

The best of Vega images of the Halley's comet nucleus was taken by the NAC camera on board the second craft at the closest encounter on 9 March 1986



Figure 2. Halley main image.

(at a distance of 8000 km) from the nucleus (see Fig. 2). Combined results of the imaging/spectral instruments revealed the shape and size of the nucleus, its anomalously low albedo even for a “dirty snowball” model (only 3–4%), the configuration of the jets, and the basic chemical components of escaping gas. The irregular shape of the nucleus could be best approximated by an ellipsoid with dimensions of $15 \text{ km} \times 8 \text{ km} \times 7 \text{ km}$.

In situ data of the space physics package included information on the intrinsic structure of the bow shock front, the product of specific “collisionless” interaction of the solar wind with plasma of cometary origin that generates hydromagnetic waves and accelerates the ions.

The various dust counters registered impacts of particles in the range from submicron to hundreds of microns. The ingenious design of the time-of-flight type mass spectrometer (shared with the Giotto spacecraft) helped to reveal several chemically different families of dust particle populations escaping from the nucleus.

The in-flight data from the various experiments on board the flyby spacecraft were complemented by a large number of remote observations both from space and from the ground; the latter were coordinated by the International Halley Watch (IHW). The IACG and its counterpart on the ground, the IHW, formed the cornerstones of a global effort to explore Halley’s comet as completely as possible during its 1985/86 encounter.

BIBLIOGRAPHY

1. Sagdeev, R.Z., J. Blamont, A.A. Galeev, et al. Vega spacecraft encounters with comet Halley. *Nature* 321: 259 (1986).
2. Sagdeev, R.Z., F. Szabo, G.A. Avanesov, et al. Television observation of comet Halley from Vega spacecraft. *Nature* 321: 262 (1986).
3. Combes, M., V.I. Moroz, J.F. Crifo, et al. Infrared sounding of comet Halley from Vega 1. *Nature* 321: 266 (1986).
4. Krasnopolsky, V.A., M. Gogoshev, G. Moreels, et al. Spectroscopic study of comet Halley by the Vega 2 three channel spectrometer. *Nature* 321: 269 (1986).
5. Moreels, G., M. Gogoshev, V.A. Krasnopolsky, et al. Near-ultraviolet and visible spectrophotometry of comet Halley from Vega 2. *Nature* 321: 271 (1986).
6. Keppler, E., V.V. Afonin, C.C. Curtis, et al. Neutral gas measurements of comet Halley from Vega 1. *Nature* 321: 273 (1986).
7. Vaisberg, O.L., V.N. Smirnov, L.S. Gorn, et al. Dust coma structure of comet Halley from Sp-1 detector measurements. *Nature* 321: 274 (1986).
8. Mazets, E.P., R.L. Aptekar, S.V. Golenetskii, et al. Comet Halley dust environment from Sp-2 detector measurements. *Nature* 321: 276 (1986).
9. Simpson, J.A., R.Z. Sagdeev, A.J. Tuzzolino, et al. Dust counter and mass analyser (DUSMA) measurements of comet Halley’s coma from Vega spacecraft. *Nature* 321: 278 (1986).
10. Kissel, J., R.Z. Sagdeev, J.L. Bertaux, et al. Composition of comet Halley dust particles from Vega observations. *Nature* 321: 280 (1986).
11. Gringauz, K.I., T.I. Gombosi, A.P. Remizov, et al. First *in situ* plasma and neutral gas measurements at comet Halley. *Nature* 321: 282 (1986).
12. Somogyi, A.J., K.I. Gringauz, K. Szego, et al. First observations of energetic particles near comet Halley. *Nature* 321: 285 (1986).

13. Riedler, W., K. Schwingenschuh, Ye. G. Yeroshenko, et al. Magnetic field observations in comet Halley's coma. *Nature* 321: 288 (1986).
14. Grard, R., A. Pedersen, J.-G. Trotignon, et al. Observations of waves and plasma in the environment of comet Halley. *Nature* 321: 290 (1986).
15. Klimov, S., S. Savin, Ya. Aleksevich, et al. Extremely low frequency plasma waves in the environment of comet Halley. *Nature* 321: 292 (1986).
16. Sagdeev, R.Z., V.M. Linkin, J.E. Blamontt and R.A. Preston. The Vega Venus balloon experiment. *Science* 231: 1407 (1986).
17. Kremnev, R.S., V.M. Linkin, A.N. Lipatov, et al. Vega balloon system and instrumentation. *Science* 231: 1408 (1986).
18. Sagdeev, R.Z., V.M. Linkin, V.V. Kerzhanovich, et al. Overview of Vega Venus balloon in situ meteorological measurements. *Science* 231: 1410 (1986).
19. Preston, R.A., C.E. Hilderbrand, G.H. Purcell, et al. Determination of Venus wind by ground-based radio tracking of the Vega balloons. *Science* 231: 1414 (1986).
20. Linkin, V.M., V.V. Kerzhanovich, A.N. Lipatov, et al. Vega balloon dynamics and vertical winds in the Venus middle cloud region. *Science* 231: 1417 (1986).
21. Linkin, V.M., V.V. Kerzhanovich, A.N. Lipatov, et al. Thermal structure of the Venus atmosphere in the middle cloud layer. *Science* 231: 1420 (1986).
22. Blamont, J.E., R.E. Young, A. Seiff, et al. Implications of the Vega balloon results for Venus atmospheric dynamics. *Science* 231: 1422 (1986).
23. Hirao, K., T. Itoh. The Planet-A encounters. *Nature* 321: 294 (1986).
24. Reinhard, R. The Giotto encounter with comet Halley. *Nature* 321: 313 (1986).
25. Munch, R.E., R.Z. Sagdeev and J.F. Jordan. Pathfinder: Accuracy improvement of comet Halley trajectory for Giotto navigation. *Nature* 321: 318 (1986).

ROALD SAGDEEV
University of Maryland
College Park, Maryland

VENUS MISSIONS

Missions

Launches of Soviet spacecraft to Venus were started in 1961 (1,2). A full list of them is presented in Table 1. It includes all Soviet launches, successful and unsuccessful, declared officially and not declared. Until 1965, the leading industrial organization responsible for lunar and planetary spacecraft was OKB-1 (Osoboe Konstruktorskoe Byuro No. 1, in translation "Special Design Bureau No. 1") managed by Chief Designer S.P. Korolev. It developed designs, manufactured the spacecraft, made tests, and controlled spacecraft in flight. Simultaneously, it had a very extensive program of manned Earth orbiting missions. It became more and more difficult to be responsible for everything, and in 1965 the lunar and planetary projects were transferred from OKB-1 to another facility, the Design Bureau and Plant named after S.A. Lavochkin. The Design Bureau ("KB") was managed by Chief Designer G.N. Babakin (1914–1971). Later (1974), this facility was renamed Nauchno-Proizvodstvennoe Obyedinenie imeni S.A. Lavochkina (NPOL), in translation Research-Industrial Association S.A.

Table 1. Soviet Missions to Venus

[Venera] 1VA No.1	probe	4 Feb 1961	Molniya
Identical to Venera 1. Failed to depart from low Earth orbit due to fourth-stage failure (2). Registered as Sputnik 7 in the United States. Carried science instruments and a pennant in a “landing apparatus,” which was an atmospheric entry probe expected to survive landing on the surface.			
Venera 1	probe	12 Feb 1961	Molniya
Communication failed in transit. Attitude control and radio system failures (2).			
[Venera] 2MV-1 No.1	atm/surf probe	25 Aug 1962	Molniya
A new design of Venera spacecraft. Carrier vehicle with a detachable entry probe. Probe expected to survive landing and carried a pennant. Failed to depart from low Earth orbit. Fourth-stage engine orientation system failed (2). Registered as Sputnik 23 in the United States.			
[Venera] 2MV-1 No.2	atm/surf probe	8 Sept 1962	Molniya
Carried a pennant in an entry probe. Failed to depart from low Earth orbit. Fourth stage failed (2). Registered as Sputnik 24 in the United States.			
[Venera] 2MV-2 No.1	flyby	12 Sept 1962	Molniya
Photo-flyby mission. Failed to depart from low Earth orbit. Third stage failed after 531 s (2). Registered as Sputnik 25 in the United States.			
[Venera] 3MV-1A No.4A	test mission	19 Feb 1964	Molniya
Test launch of new spacecraft with launch vehicle. Launch failure. Third-stage engine failure (2).			
[Venera] Cosmos 27	atm/surf probe	27 Mar 1964	Molniya
Failed to depart from low Earth orbit. Fourth-stage engine did not ignite due to power supply failure (2).			
[Venera] Zond 1	atm/surf probe	2 Apr 1964	Molniya
Spacecraft flew toward Venus, but communications failed in transit after 2 months. Leak in pressurized “orbital” section caused loss of thermal control and failure of transmitters. Radio communications conducted through probe (2).			
Venera 2	flyby	12 Nov 1965	Molniya
Flew by Venus at 23,950 km on 27 Feb 1966. Communications failed during Venus flyby due to failure of thermal control system. No data returned (2).			
Venera 3	atm/surf probe	16 Nov 1965	Molniya
Communications failed 17 days before arrival at Venus. Intended delivery of science instruments and a pennant with the USSR state emblem at arrival on 1 March 1966. First spacecraft to impact another planet (2).			
[Venera] Cosmos 96	flyby	23 Nov 1965	Molniya
Failed to depart from low Earth orbit. Third stage terminated improperly, fourth stage did not ignite due to unstable flight, and spacecraft separated with large disturbances (2).			

Table 1. (Continued)

Venera 4	atm/surf probe	12 Jun 1967	Molniya
Beginning with Venera 4, the full responsibility for planetary missions was transferred from OKB-1 to NPOL. First successful planetary atmospheric probe on 18 Oct 1967. Entered at 19°N 38°E on the night side and transmitted for 94 minutes. Measured temperature, pressure, wind velocity, and CO ₂ , N ₂ , and H ₂ O content over 25–55 km on night side of planet. Showed that the atmosphere is 90–95% CO ₂ , and measured a temperature of 535°K before being crushed at 25 km. Detected no N ₂ . Carrier vehicle included plasma and UV radiation experiments (3).			
[Venera] Cosmos 167	atm/surf probe	17 Jun 1967	Molniya
Same design and science as Venera 4. Failed to depart from low Earth orbit.			
Venera 5	atm/surf probe	5 Jan 1969	Molniya
Successful atmospheric probe; entered night side on 16 May 1969 at 3°S 18°E. Probe measured temperature, pressure, wind velocity, and CO ₂ , N ₂ , and H ₂ O content over 25–55 km on night side of planet. Transmitted for 53 minutes in the atmosphere. Lander crushed at 18 km. Flyby science same as Venera 4 (3).			
Venera 6	atm/surf probe	10 Jan 1969	Molniya
Same design and science as Venera 5. Successful atmospheric probe; entered night side 17 May 1969 at 5°S 23°E. Transmitted for 51 minutes in the atmosphere. Lander crushed at 18 km. Venera 5 and 6 found that the atmosphere is 93–97% CO ₂ , 2–5% N ₂ , and less than 4% O ₂ (3).			
Venera 7	atm/surf probe	17 Aug 1970	Molniya
First successful planetary lander on 15 Dec 1970. Landed on night side at 5°S 351°E and transmitted for 23 minutes from the surface. Measured a temperature of 747 K on the surface. No pressure data transmitted due to a failure in the data acquisition system. The Venera 7 probe was the first to survive atmospheric heat and pressure and reach the surface (3).			
[Venera] Cosmos 359	atm/surf probe	22 Aug 1970	Molniya
Same design and science as Venera 7. Failed to depart from low Earth orbit.			
Venera 8	atm/surf probe	27 Mar 1972	Molniya
Landed 22 Jul 1972 on day side near terminator at 10°S 335°E. Returned atmospheric temperature, pressure, wind speed, composition, and light levels during descent. Transmitted data for 50 min on the surface and reported a K–U–Th gamma-ray surface composition analysis. Measured a temperature of 743°K and a pressure of 93 bar at the surface (3).			
[Venera] Cosmos 482	atm/surf probe	31 Mar 1972	Molniya
Same design and science as Venera 8. Failed to depart from low Earth orbit. Fourth stage misfired.			

Table 1. (Continued)

Venera 9	orbiter/lander	8 Jun 1975	Proton-D
New design of heavy Venera spacecraft using the Proton launcher. Dispatched successful lander and orbited Venus 22 Oct 1975. The first Venus orbiter, first picture from the surface of another planet, and first use of orbiter as relay for a planetary probe. Descent probe landed on day side at 32°N 291°E and communicated through the orbiter for 53 minutes. Lander measured atmospheric composition, structure, and photometry on descent and obtained B/W images and K–U–Th gamma-ray analysis on the surface. Orbiter returned imagery, IR-radiometry, spectrometry, photopolarimetry, radio occultation, and plasma data (4,5).			
Venera 10	orbiter/lander	14 Jun 1975	Proton-D
Same design and science as Venera 9. Dispatched successful lander on day side at 16°N 291°E and orbited Venus 25 Oct 1975. Venera 9 and 10 found the lower boundary of the clouds at 49 km and three distinct cloud layers at altitudes of 57–70 km, 52–57 km, and 49–52 km. Both orbiters ceased operations in March 1976 (4,5).			
Venera 11	flyby/lander	9 Sep 1978	Proton-D
Landed 25 Dec 1978 on the day side at 14°S 299°E. Measured atmospheric temperature, pressure, wind velocity, spectra of short wavelength radiation, chemical and isotope composition, aerosols, and thunderstorm activity. Surface imaging and XRF sapling systems failed. Contact lost after 95 minutes on the surface. Flyby spacecraft carried UV spectrometer, plasma instruments, and lander relay communications (6).			
Venera 12	flyby/lander	14 Sep 1978	Proton-D
Same design and science as Venera 11. Landed 21 Dec 1978 on the day side at 7°S 294°E. Same science as Venera 11 and included cloud particle composition. Surface imaging and XRF sapling systems also failed. Continued transmitting data for 110 minutes until flyby spacecraft went below the horizon (6).			
Venera 13	flyby/lander	30 Oct 1981	Proton-D
Landed 1 Mar 1982 on the day side at 7.5°S 303.0°E. Conducted atmospheric and cloud science and both BW and color imagery of the surface, as well as XRF analysis of the surface material. Contact with lander lost after 127 minutes (8).			
Venera 14	flyby/lander	4 Nov 1981	Proton-D
Same design and science as Venera 13. Landed 5 Mar 1982 on the day side at 13.4°S 310.2°E. Contact with lander lost after 63 minutes (8).			
Venera 15	orbiter	2 Jun 1983	Proton-D
Entered Venus orbit 10 Oct. Radar mapper covered the planet from 30°N to North Pole at 1–2 km resolution. The middle atmosphere and clouds were examined by IR spectrometry (9,10).			
Venera 16	orbiter	7 Jun 1983	Proton-D
Same design and science as Venera 15. Entered Venus orbit 14 Oct. Radar mapper with same coverage and resolution as Venera 15. IR instrument failed (9,10).			

Table 1. (Continued)

Vega 1	flyby/lander/balloon	15 Dec 1984	Proton-D
Venus flyby 11 June 1985 using a gravity assist maneuver to redirect the spacecraft to Halley's comet. Deployed an entry vehicle with a balloon and lander on the night side of the planet at 8.1°N 176.7°E. Conducted atmospheric science on descent. Balloon released on descent and floated for 48 hours measuring downdrafts of 1 m/s and average horizontal wind of 69 m/s. Drifted approx 10,000 km at about 54 km altitude. Lander soil analysis was made by the gamma spectrometer; the drilling device to provide a sample for the X-ray fluorescence spectrometer started to work in the atmosphere and thus failed to get a sample. Spacecraft bus flew by Venus and continued on to flyby Halley's Comet at 8890 km distance on 6 Mar 1986 (11,12).			
Vega 2	flyby/lander/balloon	21 Dec 1984	Proton-D
Same design and science as Vega-1. Deployed balloon and lander in the night-side Venus atmosphere at 7.2°S 179.4°E on 15 Jun 1985 flyby with similar results. Same measurements as Vega 1. Sample acquisition with XRF and gamma-ray analyses on the surface were both successful. Spacecraft bus continued on to flyby Halley's Comet at 8030 km distance on 9 Mar 1986 (11,12).			

Lavochkin; the brief name "Lavochkin Association" is also used sometimes unofficially. For simplicity, we will use the designation "NPOL" for all periods covered below.

Almost all Soviet missions to Venus were named "Venera" because Venera is the name of this planet in Russian. The only exception was the last one, it was "Vega," not "Venera" (see below). OKB-1 made eleven attempts to launch Venera spacecraft from February 1961 to November 1965, but all of them were unsuccessful; either the launcher failed, or something failed in the spacecraft systems on the way to the planet. Venera 1 got on a trajectory to the planet, but communication was lost in the middle of the journey. Venera 3 reached the planet, but communications failed 17 days earlier. Some brief information about other failed missions is given in Table 1. In reality, they were not vain efforts. Very new techniques were created, and time was necessary for step by step developments, tests, and improvements. Meanwhile, the U.S. Mariner 1 went off course during launch in July 1962. A month later Mariner 2 was launched successfully and after a 3.5-month flight, flew by Venus and scanned the planet with infrared and microwave radiometers. That was the first successful planetary mission.

The first success in space studies of Venus was achieved by the Soviets in 1967. Venera 4 reached the planet and fulfilled its goals. It was separated from the bus and entered the atmosphere, slowing as it descended. The parachute opened, and measurements of the atmospheric properties were taken down to an altitude 25 km above the surface. The probe was destroyed there due to high density and/or high temperature. Venera 4 was the first successful planetary atmospheric probe. By this flight scheme, the bus from which the probe is separated enters the atmosphere and burns up. Separation was made several days before arrival. Data from such probes were transmitted to Earth directly by a low-gain antenna with a slow rate of 1 bit/s.

Accidentally or not, this success coincided with the above mentioned transfer of work from OKB-1 to NPOL. The technical documentation for Venera 4 had been prepared mainly in OKB-1, although NPOL designers did some updating. Of course, the Venera 4 design embodied the entire experience of OKB-1 in the development of planetary spacecraft. However, it was reinforced by their own NPOL know-how in design, manufacturing, and tests of rocket/aviation technology. After Venera 4, all Soviet planetary spacecraft were built by NPOL with its own team in cooperation with many other industrial firms.

The U.S. flyby probe, Mariner 5, arrived at Venus 1 day after Venera 4 and conducted some remote studies of the atmosphere. Then for about 10 years, NASA did nothing to explore Venus besides the Mariner 10 flyby on its way to Mercury. In contrast, the Soviet Union at that time was sending missions to Venus regularly, in the beginning using each astronomical window (about a 1.5-year gap) and later each second window (about 3 years). Other spacecraft that had increasingly better system environment protection followed Venera 4. Venera 7 (1970) made the first soft landing and transmitted signals directly from the surface. However, something failed in the data acquisition system, and only temperature data were transmitted, although they were much more detailed than those taken on earlier missions.

All probes from Venera 4 to Venera 7 landed on the night side of the planet due to navigational requirements. However, a day-side landing was very desirable for understanding the Venusian climate. This would provide the possibility of seeing if solar light reaches the surface of the planet or not. For this reason, Venera 8 (1972) was intentionally landed on the illuminated part of the planet, although very near the terminator. Measurements showed that some solar energy flux actually reaches the surface, as required by the greenhouse effect concept.

The first set of missions to Venus, up to Venera 8, was conducted with the Molniya launchers (the R7 rocket plus fourth-stage L). These missions and their results are described in Ref. 3. Since 1975, NPOL used the more powerful rocket Proton (plus fourth-stage D). A new generation of Venera spacecraft was born due to this change, larger than the previous, with the conversion of the bus to the orbiter or flyby module. In both cases, the bus was used for the lander data relay. In Venera 9 and 10 missions, the first Venus orbiters were created, and new landers descended with greater success than previous ones. These landers (Fig. 1) transmitted to Earth the first panoramas of the Venusian surface. After this, there were six more successful Soviet missions (Venera 11, 12; Venera 13, 14; Vega 1, 2) with landers (but without orbiters) and two missions with orbiters (but without landers), Venera-15 and 16 with the synthetic aperture radar (SAR). The results of these latest missions are described in Refs. 4–13.

All Soviet planetary missions (except the last, Mars 96) were duplicated: two identical spacecraft were always launched. In Soviet conditions, it cost only 15–20% more than a single mission, but such duplication provided a significant increase in the overall mission reliability.

Soviet Venera landers were unique. The new generation landers survived on the surface not less than two hours providing panoramic imaging (B/W and color) and measurements of the surface material composition by X-ray fluorescence (XRF) spectrometry. The XRF experiment was extremely difficult because

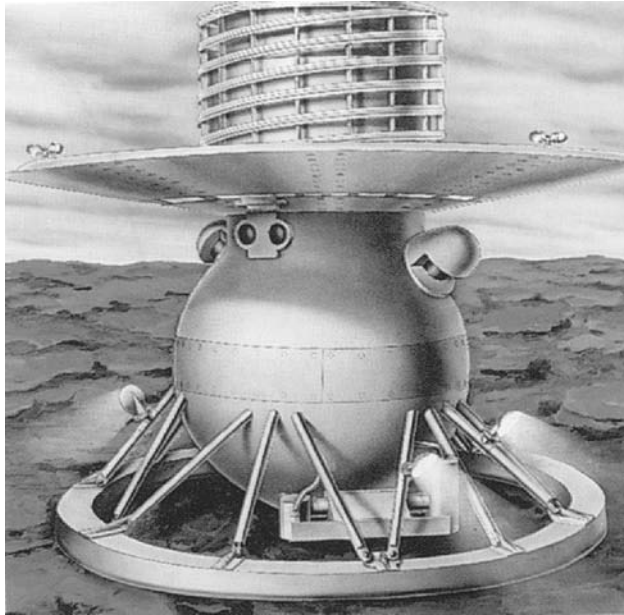


Figure 1. Venera 9 and 10 probe, general view. An artist's concept: the probe after landing (22 and 25 October 1975). Full mass about 1560 kg. Diameter 2.4 m. They transmitted scientific information (including panoramas) from the surface of Venus for 53 and 65 min, respectively.

of the necessity of bringing a sample into the interior part of the lander. In all cases, communication between the lander and the bus were interrupted only by geometry, not due to limited lander lifetime.

The latest Soviet missions to Venus, Vega 1 and Vega 2 (1984) were very ambitious. Together with “classical” landers, balloons were targeted at the Venusian atmosphere, and after that the buses were redirected to Halley’s comet, becoming cometary probes. The Vega mission was prepared and conducted with broad international cooperation. Scientific and technical groups from eight countries participated in developing the science payload. Twenty radio astronomical observatories observed drifts of balloons in the atmosphere using VLBI and Doppler methods. This network was organized by Centre National d’Etudes Spaciales (CNES) which was especially interested in the balloon part of the Vega missions.

After Vega 1 and 2, no more Soviet missions were sent to Venus. Possibly it was not the right decision. In reality, exploration of Venus was something like a Soviet “ecological niche” within the world of space science. For economic reasons, the Soviet Union could not compete with the United States in every field of space research. M.V. Keldysh mentioned that in such conditions a reasonable strategy must be to concentrate efforts in some narrow selected directions. Venusian exploration was just such a direction in space science.

Several books were dedicated to the analysis of numerous results of Soviet and U.S. missions to Venus (14–20). The Soviet scientific input will be outlined very briefly below.

Results of Studies: The Atmosphere and Its Interaction with the Solar Wind

Vertical Structure of the Atmosphere from the Surface to 100 km. A lot of *in situ* measurements were made by Venera probes using temperature and pressure sensors at altitudes from about 60 km to the surface. It was shown that the temperature and pressure profiles are nearly the same on the day and night sides of Venus at these altitudes. Venera and Pioneer results obtained before 1979 were combined in the COSPAR Venus Reference Atmosphere, VIRA (16). At zero reference level (corresponding to a radial distance 6052 km from the center of mass of the planet), the temperature is 735°K, and the pressure is 92 bar according to VIRA. The temperature and pressure decline with altitude to values of about 260°K and 0.2 bar at 60 km. An important input was provided by Vega 2, the latest of all Venus probes (see Table 1). For the first time, very precise measurements of T , P profiles were made down to the surface (12). The temperature profile is presented in Fig. 2 These T , P measurements were better than those on previous Soviet Venera landers. Temperature sensors on the U.S. Pioneer probes had nearly the same precision as those on Vega but failed below 11 km. The analysis of the Vega 2 temperature profile at low altitudes shows the presence of convection in lower layers of the atmosphere. This region of convective instability probably extends up to altitudes of 25–30 km. The atmosphere is stable between 30 and 50 km, but above 50 km (and up to 55 km) becomes unstable again. Consequently, the atmosphere of Venus (in contrast to the terrestrial one) has two convective zones separated by a wide stable interval with a subadiabatic lapse rate. The Vega 1 and Vega 2 balloons (13) drifted at altitudes of 52–53 km just inside the upper convective zone and confirmed the presence of pronounced turbulence there.

The vertical profiles of the atmosphere between 60 and 90 km were measured by several means, including accelerometers on descent probes and infrared

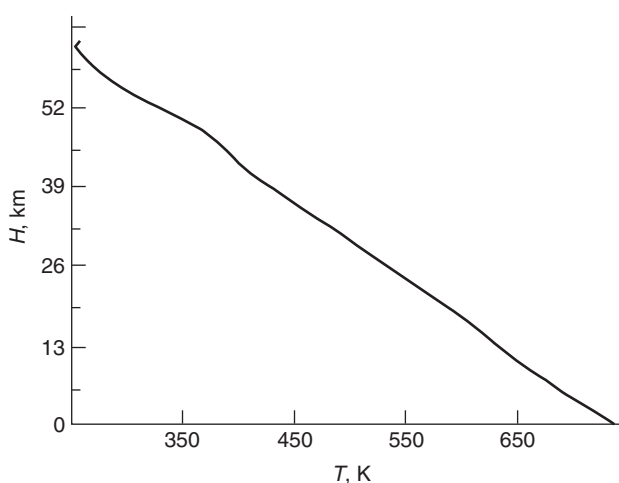


Figure 2. Vega 2 probe, the measured vertical profile of atmospheric temperature.

spectrometry of the thermal radiation emitted to space (Venera 15). The analysis of orbiter (Venera 9, 10, 15, 16) radio occultation data gave a set of profiles between 40 and 90 km. Observations with an IR Fourier spectrometer FS-1/4 on board the Venera 15 orbiter (9) provided about 2000 spectra in the range from 6 to 35 μm with a spectral resolution of 5–7 cm^{-1} and spatial resolution of about 200 km. These spectra contain rich information about the atmospheric temperatures, aerosol, and some minor constituents (H_2O and SO_2). The altitude temperature profiles were found from the spectral shape of the strong CO_2 band centered at 15 μm . The temperature at the altitude of 90 km is about 170°K. The shape of the altitude profile depends on latitude. In equatorial and midlatitudes ($<45^\circ$), there is a monotonic decrease with altitude. In latitudes $>45^\circ$, temperature inversions can be seen ordinarily with a minimum near 60–65 km and a maximum near 70–75 km. These structures are generated by the dynamic transfer of heat. Solar time and local variations were also observed.

Winds. There are very strong zonal winds in the atmosphere of Venus. The principal mode of global circulation in the Venusian atmosphere is retrograde superrotation; the entire atmosphere below 85 km moves in the same direction as the solid planet itself, but at much greater speed. Maybe only the lowest layers move in the opposite direction, but there are no measurements of wind direction in that part of the atmosphere. The first estimates of wind speeds below 60 km were obtained by Doppler tracking of Venera 4. Similar measurements were made on all later probes. On Vega 1 and 2, observations of balloon motions were added: about 70 m/s at a nearly fixed altitude of 54 km. Wind speed decreases below this altitude down to several m/s at 10 km. Wind velocities above 40 km were derived also from the thermal profiles obtained by Venera 9, 10, 15, 16 radio occultations using cyclostrophic balance equations. Venera 15 thermal profiles obtained by IR spectrometry were used even more effectively for wind retrievals. They show wind speeds up to 130 m/s at altitudes of 70–75 km with two pronounced daily maxima, as expected because of thermal tides. Close to the surface (at an altitude of 1.3 m) on Venera 9 and 10 landers, a wind speed of about 1 m/s was measured by anemometers. It is not possible to say anything about the direction of this surface wind.

Chemical and Isotopic Composition of Atmospheric Gases. The first direct measurements of atmospheric chemical composition were made with simple chemical sensors on earlier Venera missions. They showed that CO_2 and N_2 are the main constituents (97% and 3%, taking into account also some latest qualifications). Much more sophisticated instruments have worked on later probes (from Venera 11 to 14), namely, mass spectrometers and gas chromatographs. The results of mass spectrometry confirmed a strong anomaly of the argon isotope ratio ($^{36}\text{Ar}/^{40}\text{Ar}$) discovered several months earlier by the U.S. Pioneer large probe. This ratio on Venus is about 1 instead of 0.003 as in the terrestrial atmosphere. It was an important observation because of the different nature of these isotopes; one of them (^{36}Ar) is so-called primary, the second (^{40}Ar) is radiogenic. It was found that all primary (nonradiogenic) isotopes of noble gases are much more abundant on Venus than on Earth. Ratios of nonradiogenic isotopes of noble gases in some cases also showed some small differences from terrestrial ratios. It was found (on Venera 13 and 14) that the $\text{Ne}^{20}/\text{Ne}^{22}$ ratio is 12.2. This is larger than that on Earth (10.1) and less than that in the solar wind

(13.7). This means that solar wind implantation may influence the evolution of the atmosphere of Venus.

Gas chromatographs measured the abundance of several gases. Among them are SO₂ and CO (130 and 28 ppm, respectively, at 42 km in altitude, "Venera 12"). Later, the UV spectrometer on Vega probes did find that the SO₂ mixing ratio is 20–25 ppm at 12 km, 38 ppm at 22 km, and 125–140 ppm at 42 km (the last in near agreement with Venera 12). SO₂ abundance in the upper clouds (~70 km) is much less and varies very strongly with latitude: 0.03 ppm at the low and middle latitudes and 0.100 to 1 ppm in the North Polar region according to IR spectra obtained by the Venera 15 Fourier spectrometer. There are significant place-to-place and time variations of SO₂ abundance in the upper clouds.

H₂O abundance in the lower atmosphere was estimated many times using different kinds of instruments including chemical sensors, gas chromatographs, and optical spectrophotometers (Fig. 3). There was a big discrepancy among the results obtained in these measurements, from 30 to 1000 ppm, approximately. It seems now that only optical spectrophotometers provided a realistic estimate. According to optical measurements, the H₂O column abundance is about 1 g/cm², and the mixing ratio is 30–40 ppm between the surface and the main cloud deck. Low resolution spectrophotometers were installed on Venera 11, 13, and 14 to measure the spectrum of the solar radiation diffusing through the atmosphere in the range of 0.4–1.2 μm. These spectra contain several pronounced H₂O and CO₂ absorption bands (Fig. 3). In spite of the relatively low abundance of H₂O, this gas provides a substantive input to the overall infrared opacity for the upwelling thermal ($\lambda > 2 \mu\text{m}$) radiation in the atmosphere of Venus. H₂O abundance within the clouds is even lower. About 5–10 ppm (with some latitudinal, place-to-place, and daily variations) was found in the upper clouds from analysis of spectra obtained with Venera 15 IR spectrometry.

Clouds and Hazes. Clouds cover the whole planet Venus, making its surface invisible from outside. It was known from ground-based observations that their upper boundary is at an altitude of about 65–70 km above the surface. Venera 9 and 10 probes (1975) discovered their lower boundary at an altitude of about 49–50 km. It was found independently by two experiments, nephelometer and photometer. They showed also that the main cloud deck (located between these two boundaries) consists of three layers of different optical density: the lower (the most dense), middle (most transparent), and upper clouds. The atmosphere above and below the main cloud deck is also not free from aerosols: there is an upper haze above the upper clouds and a lower haze below the lower clouds. In the main cloud deck, the typical sizes of particles are of the order of 1–2 μm according to measurements with the nephelometer. The size spectrum is wide and varies with altitude. The average optical thickness of the main cloud deck is about 30. Direct solar radiation does not penetrate through the clouds, but there is a high enough flux of scattered solar radiation so that cloud particles consist of almost transparent material. The vertical structure of the upper clouds was studied in detail by IR spectra obtained by Venera 15. They showed that the aerosol scale height in the low and middle latitudes is large, about 5–6 km, but is much smaller (down to 1–2 km) in high latitudes. An important fact derived from these spectra is that their continuum (outside of CO₂ and H₂O bands) is in fair agreement with micron

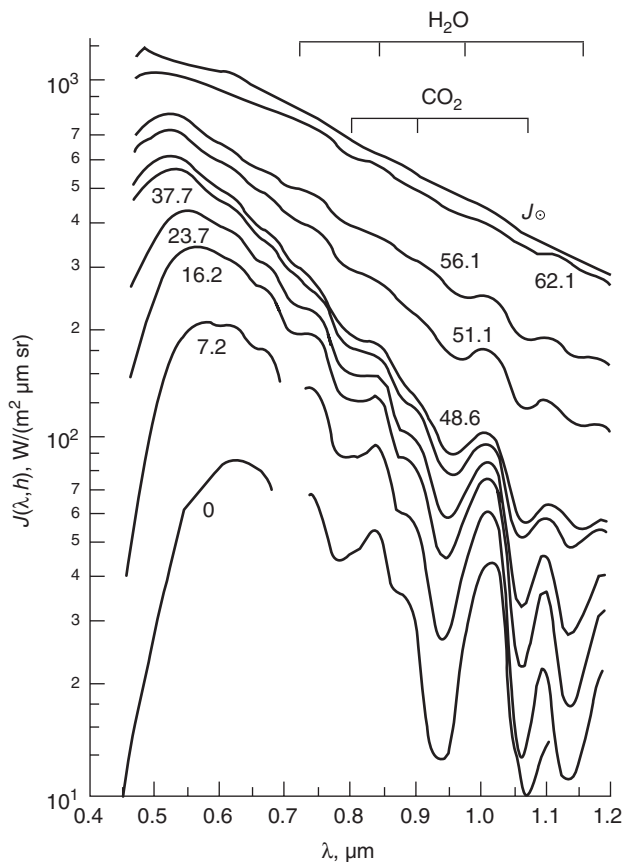


Figure 3. Venera 11: Examples of spectra of solar light penetrating the lower atmosphere of Venus due to atmospheric scattering. Numbers near curves are altitudes above the surface (in km). The overall shape of spectra and intensities are defined by extinction in clouds and atmospheric gas. CO_2 and H_2O absorption bands are easily visible.

size particles that consist of sulfuric acid (at a concentration of about 75%). However this does not mean that H_2SO_4 is the only chemical component of the particulate matter in clouds at all altitudes. Middle and lower clouds may include particles of other chemicals. The strongest evidence of this was obtained by one of the experiments of the Vega probes; cloud particles were collected, and their elemental composition was estimated by a simplified X-ray fluorescent method. Sulfur was identified everywhere in the main cloud deck, but chlorine and phosphorus were also found in the middle and especially in the lower clouds. Moreover, phosphorus, not sulfur dominated here.

Solar and Thermal Fluxes. The photometer installed on Venera 8 discovered that part of the solar light penetrating through the atmosphere reaches the surface after multiple scattering in the clouds and gas. More sophisticated (multiband) photometers on Venera 9, 10, 11, 12 and spectrophotometers (with continuous spectral scanning) on Venera 11, 13, 14 provided much more information about light fluxes from different directions and at all altitudes from about 60 km

down to the surface. They showed, in particular, that the illumination of the surface (spectrally integrated) is about 3–5% of the solar illumination at the upper boundary of the clouds. The surface albedo is low (about 0.1), so that the incoming solar flux is absorbed by the surface and warms it. This minor energy flux (together with the high opacity of the atmosphere to infrared thermal radiation) may be enough to support the high temperature of the surface and the lower atmosphere due to the greenhouse effect.

The thermal radiation within the atmosphere was not measured by Soviet probes, but there were measurements by orbiters of outgoing thermal radiation from the planet. The first time, it was measured by Venera-9 and 10 by a very simple single filter radiometer, and later on Venera 15 by the IR Fourier spectrometer. The first of these sets of measurements was made only at low latitudes, but the second covered a wide latitude range from the equatorial to polar regions. This second confirmed the strong peculiarity of the thermal radiation field of Venus. The radiance is higher in polar regions than in low latitudes, as was discovered some years earlier by the Pioneer orbiter.

Airglow. Night-side emission in the range 4000–7000 Å was discovered by high sensitivity spectrometers of the Venera 9 and 10 orbiters. The observed spectrum consisted of seven peaks that were later identified as a superposition of several bands of molecular oxygen, O₂. These bands can be emitted only if O₂ is a small constituent. For this reason, they dominate on Venus, but are absent in the terrestrial night emission spectrum. Possibly, the “ashen light” of Venus that is observed sometimes from Earth may be explained by this emission. It is generated at an altitude of 100 km. Far UV emission in the lines of H, He, O, and O⁺ was measured by filter photometers of Venera 9, 10 and spectrometers of Venera 11, 12 flyby modules. Abundances of these atomic constituents and the upper atmosphere temperature were estimated by analysis of these data.

Electrical Activity in the Lower Atmosphere. The experiment “Groza” (“Thunderstorm”) was conducted on Venera 11, 12, 13, and 14 for observations of low frequency (LF) electromagnetic emission in the range 10–80 kHz. Sporadic (impulsive) radiation was actually observed whose general characteristics were similar to the LF emission of terrestrial thunderstorms. It was supposed originally that this emission is generated in the clouds. Another hypothesis assumes that a possible source may be located much lower and even connected with some volcanic activity. An optical flash on the night side of Venus (lightning) was observed once by the spectrometer of Venera 9.

Ionosphere. Multiple radio occultation of orbiters Venera 9 and 10 on two frequencies (wavelengths 32 and 8 cm) provided the first evidence of strong variability of the electron density (n_e) profile depending on solar zenith angle and conditions of interaction with the solar wind. During daytime, the main maximum (with n_e about $4 \times 10^5 \text{ cm}^{-3}$) was located at an altitude of about 150 km, and the ionosphere goes up to 300–800 km (ionopause level). The night ionosphere had its main peak (about 10^4 cm^{-3}) at 130–140 km, and there were no electrons above 170–200 km. Sometimes, two peaks were visible in the night profiles. Additional measurements were made by radio occultation of Venera 15 and 16.

Solar Wind Interaction with the Atmosphere. This was investigated by two plasma spectrometers and a magnetometer onboard the Venera 9 and 10

orbiters. They show the constant presence of a bow shock, which heats and compresses the solar wind flow. Venus has no intrinsic magnetic field, but a so-called magnetosheath is generated on the internal side of the bow shock. The ionized flow of the magnetosheath can interact directly with the neutral atmosphere due to charge exchange and photoionization. This adds mass to the solar wind because the upper atmosphere consists mainly of oxygen. The Venusian ionosphere is not protected from interaction with this flow due to an absence of an intrinsic magnetic field. The planetary ion flow on the night side forms a planetary plasma tail. The night ionosphere is supported by the plasma flow from the day side and also by electron precipitations. Findings of Venera 9 and 10 in studies of the solar wind interaction with the ionosphere of Venus were developed later by experiments on board Pioneer Orbiter.

Results of Studies: The Surface

Surface Morphology and Geology. These have been studied by Soviet missions to Venus on two different scales. In 1983–1984, Venera 15 and 16 orbiters provided side-looking radar images of the northern 21% of the planet surface at 1–2 km spatial resolution. An example is presented in Fig. 4. These were the first images on which landforms larger than a few kilometers across were seen, and thus local and regional geological structures could be easily identified. Simultaneously, the topography of that part of the planet was measured with 5×50 km footprints and a 30-m rms error of the relative altitudes. Even earlier (in 1975 and 1982), Venera 9, 10, 13, and 14 landers sent to Earth TV panoramas (see Fig. 5) of the landing sites showing centimeter-scale details of the surface in vicinities closest to the landers and of progressively larger landforms at a distance.

Analysis of the Venera 15–16 data allowed us to understand the key features of Venusian geology, later confirmed by the U.S. Magellan mission results. It showed that the part of Venus studied was dominated by vast lava plains. Besides plains, more than 30 large (>100 km in diameter), almost 1000 intermediate (20–100 km), and more than 20,000 small (<20 km) volcanic constructs have been identified. Within the plains, “islands” and “continents” of high-standing blocks of specific terrain named “tessera” were seen. The tessera are morphologically rough due to multiple tectonic deformations predating, in most cases, the emplacement of volcanic plains. Volcanic-tectonic features, named “coronae,” circular to ovoidal and hundreds of kilometers across, were also observed within the plains. They are specific to Venus and are not observed on other planetary bodies. About 150 impact craters have been identified in the Venera 15–16 images. Their area density showed that the observed ensemble of volcanic and tectonic features and terrains had a mean surface age of about 0.5 to 1 billion years. Excellent preservation of the observed features for this long time provided evidence that surface processes such as wind erosion and deposition had very low mean rates on Venus. The observed area distribution of the features and terrains, including close to random spatial distribution of impact craters, led us to conclude that the plate-tectonic style of planet geodynamics typical of Earth was not the case in the morphologically identifiable (<1 billion

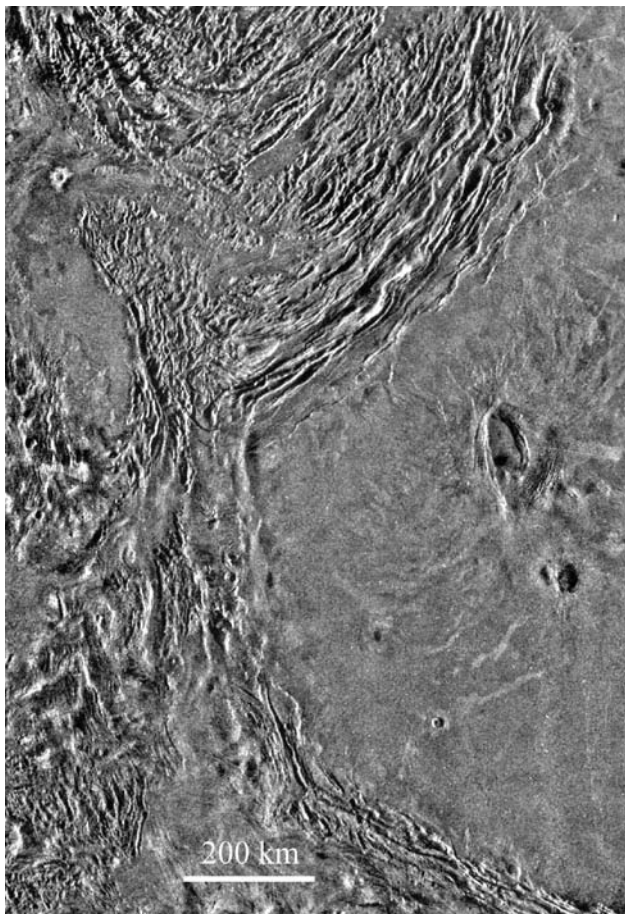


Figure 4. Venera 15: An example of the radar images showing the western part of the volcanic plateau Lakshmi and tectonic mountains Akna Montes.

years) part of the geologic history of Venus, at least for the part of the planet studied.

TV panoramas taken by Venera 9, 10, 13, and 14 spacecraft that landed thousands kilometers apart from each other showed strikingly similar small-scale morphology: bedrock consisting of sequences of close to horizontal, numer-



Figure 5. Venera 13: TV panorama showing outcrops of bedded material and soil.

ous, centimeter-thick beds, forming local highs, and structureless soil in local depressions. At the Venera 9 site, which is on a steep ($\sim 30^\circ$) slope, the bedrock is seen in the form of platy rock fragments that are part of the on-slope talus. A clod of soil several centimeters across was thrown onto the supporting ring at the landing of Venera 13. Several sequential TV pictures showed clearly that the clod was gradually removed by wind during the approximately one-hour observation time.

Surface Albedo and Color. Photometric analysis of the Venera 9, 10, 13, and 14 TV panoramas showed that the albedo of the soil is about 0.03–0.04, and the albedo of the bedrock varies from 0.05 to 0.12. Surface color was not an easy thing to determine even when Venera 13 and 14 sent color TV panoramas back to Earth because the sunlight reaching the planet surface is not neutrally white but prominently orange. The most reliable data on surface color were received by photometers of the Venera 9 and 10 probes. It was concluded that only basalts are similar to the surface of Venus in regard to spectral characteristics.

Chemical Composition of the Surface Material. This was determined by two methods, gamma-ray spectrometry and X-ray fluorescence spectrometry. The first was used by Venera 8, 9, 10, and Vega 1 and 2. The gamma-spectrometer was inside the pressure- and temperature-protected lander compartment and measured spectra of gamma irradiation of potassium, uranium, and thorium of the surface material that penetrated through the compartment walls. The measurements showed that the contents of these three radioactive elements in the surface materials of the Venera 9, 10 and Vega 1 and 2 sites were the same as in terrestrial tholeiite basalts. In the case of Venera 8, the contents of these elements found were significantly higher, reaching the level typical for such terrestrial rocks as alkaline basalt, monzonite, or syenite. A second technique was used by Venera 13, 14, and Vega 2. The landers had a drilling device and a system of delivering the acquired surface material sample to the inside of the protected lander compartment, where the measurements were done. Contents of Si, Ti, Al, Fe, Mn, Mg, Ca, K, S, and Cl were determined. In the surface materials of the Venera 14 and Vega 2 sites, the contents of the listed elements found were the same as in terrestrial tholeiite basalts. In the case of Venera 13, the content found corresponds to alkaline (leucite) basalt.

Physicomechanical Characteristics and Electric Resistivity of the Surface Material. The measurements started with attempts to measure the density of the surface material with the help of a gamma-ray densitometer at the Venera 9 and 10 landing sites. In the case of Venera 9, the attempt was unsuccessful; the densitometer sensor unit stood upon two rock fragments standing apart with a gap in between so that the sensor was not in the required contact with the solid surface. In the case of Venera 10, the sensor unit stood upon a bed rock outcrop, although the unit inclination to the outcrop surface was not exactly determined, which left the possibility of an equivocal estimate. The preferred estimate of the bulk density of the rock is $2.8 \pm 0.1 \text{ g cm}^{-3}$, but the alternative estimate of $\sim 1.5 \times \text{g cm}^{-3}$ is also possible. Then, estimates of bulk densities and bearing capacities of the surface material were done at the Venera 13 and 14 sites from dynamic loading data (when the lander impacted the surface) and using the trellis girder. These results showed that the bedrock material at these sites had

physicomechanical properties similar to those of packed sintered sand, namely, a density $1.4\text{--}1.5\text{ g cm}^{-3}$ and bearing capacity of $4\text{--}10\text{ kg cm}^{-2}$. The soil is a very weak and porous material with a density from $1.15\text{--}1.2\text{ g cm}^{-3}$ and a bearing strength $\sim 2\text{ kg cm}^{-2}$. The electric resistivity of the surface material was unexpectedly low: 73 to 89 ohm m.

Concluding Comments

A lot of valuable information about Venus was obtained by the Soviet (and also U.S.) missions to this planet. Most of the data (although still not all) has been carefully analyzed. A set of new goals and new possibilities was identified by this activity. Proposals for new missions to Venus are offered from time to time by scientists from different countries; however, space agencies have not paid serious attention to this planet in recent years. Mars and small bodies dominate their programs. We think that it is not a well-balanced approach. For a better understanding of the past and future of our Earth, we need careful and systematic studies of both of its neighbor planets, Mars and Venus, not just one. The valuable technical experience obtained earlier is still available. At the same time, new approaches are visible, like use of near-infrared day emissions for deep sounding of the atmosphere from orbit, high temperature electronics for landers and balloons, etc. Possibly it is time now for a return to the exploration of Venus on a new level.

BIBLIOGRAPHY

1. Moroz, V.I., W.H. Huntress, and I.L. Shevaleyev. Planetnye ekspeditsii XX veka [Planetary missions of the XX century]. *Kosmich. Issled.* 40 (5): 1–31, 2002 (in Russian).
2. Chertok, B.E. *Rakety i liudi. Gorjachie dni holodnoi voyny* [Rockets and People. Hot Days of the Cold War]. Moskva, Mashinostroenie, 1999 (in Russian).
3. Kuzmin, A.A., and M.Ya. Marov. [Physics of the Planet Venus]. Moscow, Nauka, 1974 (in Russian).
4. (Venera 9 and 10¹) *Kosmich. Issled.*² 14 (5 and 6): 651–877 (1976) (in Russian).
5. (Venera 9 and 10) *Geologic Soc. Am. Bull.* 88: 1537–1545 (1977).
6. (Venera 11 and 12¹) *Kosmich. Issled.*² 17 (5): 829 (1979) (in Russian).
7. (Venera 13 and 14¹) *Kosmich. Issled.*² 21 (2): 147–319 (1983) (in Russian).
8. (Venera 13 and 14) *Geologic Soc. Am. Bull.* 96: 137–144 (1985).
9. (Venera 15 and 16¹) *Kosmich. Issled.*² 23 (2): 179–267 (1985) (in Russian).
10. (Venera 15 and 16¹) *J. Geophys. Res.* 91: D378–D430 (1986).
11. (Vega and Giotto, Halley comet flyby.) *Nature* 321: 259–366 (1986).
12. (Vega 1 and 2¹) *Kosmich. Issled.*² 25 (5 and 6): 643–958 (1987) (in Russian).
13. (Vega balloons¹) *Science* 321: 1441–1480 (1986).
14. Huntten, D.M., L. Colin, T.M. Donahue, and V.I. Moroz (eds). *Venus*. The University of Arizona Press, Tucson, AZ, 1983.
15. Ksanfomality, L.V. [The Planet Venus]. Moscow, Nauka, 1985 (in Russian).

¹A special issue or set of publications in a single issue.

²English translation of *Kosmicheskie Issledovaniia* is available as *Cosmic Research*.

16. Kliore, A.J., V.I. Moroz, and G.M. Keating (eds). The Venus International Reference Atmosphere. *Adv. Space Res.* 8 (11): 1–305 (1985).
17. Barsukov, V.L., and V.P. Volkov (eds). [*The Planet Venus*]. Moscow, Nauka, 1989 (in Russian).
18. Barsukov, V.L., et al. (eds). *Venus Geology, Geochemistry, and Geophysics. Research Results from the USSR*. The University of Arizona Press, Tucson, AZ, 1992.
19. Bougher, S.W., D.M. Hunten, and R.J. Phillips (eds.). *Venus II*. The University of Arizona Press, Tucson, AZ, 1997.
20. Surkov, Yu.A. [*Cosmo-Chemical Studies of the Planets and Satellites*]. Moscow, Nauka, 1985 (in Russian).

VASILY I. MOROZ
Space Research Institute
Russian Academy of Sciences
Moscow, Russia

ALEXANDER T. BASILEVSKY
Vernadsky Institute of Geochemistry
and Analytical Chemistry
Russian Academy of Sciences
Moscow, Russia

W

WEATHER SATELLITES

Introduction to Weather Satellites

The world's first meteorological satellite, a polar-orbiting satellite, was launched from Cape Canaveral on 1 April 1960. Named TIROS for Television Infrared Observation Satellite, it demonstrated the advantage of mapping Earth's cloud cover from satellite altitudes. TIROS showed clouds banded and clustered in unexpected ways. Sightings from the surface had not prepared meteorologists for the interpretation of the cloud patterns that the view from an orbiting satellite would show.

The spacecraft, the sensors, the communication links, the data, and the data uses of weather satellites today bear little resemblance to what they were at the time of that pioneering effort. In the decades that have elapsed, satellite weight has grown from about 100 kg to nearly a metric ton; vidicon cameras have given way to scanning radiometers; hand-drawn analyses of the data have been replaced by computer-generated products; what was analog has become digital; and capabilities have expanded to include atmospheric profiling, ocean sensing, and collecting and relaying environmental data recorded by remote reporting platforms. In addition, weather satellites now measure the space environment in which they operate. Just as their technology and uses have evolved, so have weather satellites proliferated internationally. Meteorological spacecraft in polar orbit and geostationary orbits are operated by the United States, the European Space Agency, Japan, India, Russia, and the People's Republic of China.

Today, U.S. environmental satellites are operated by NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) in Suitland, Maryland. NOAA's operational environmental satellite system is composed of two types of satellites: geostationary operational environmental satellites (GOES) for national, regional, short-range warning and "now-casting," and

polar-orbiting environmental satellites (POES) for global, long-term forecasting, and environmental monitoring. The GOES and POES satellites also carry search-and-rescue instruments to relay signals from aviators and mariners in distress. Both types of satellites are necessary for a complete global weather monitoring system. In addition, NOAA operates polar-orbiting satellites in the Defense Meteorological Satellite Program (DMSP). NESDIS also manages the processing and distribution of the environmental data and images that the satellites produce each day. Figure 1 shows a current GOES spacecraft, and Fig. 2 illustrates a POES space vehicle.

NASA's Goddard Space Flight Center (GSFC) in Greenbelt, Maryland, is responsible for the procurement, development, launch services, and verification testing of the spacecraft, instruments, and unique ground equipment. Following deployment of the spacecraft from the launch vehicle, GSFC is responsible for the mission-operation phase leading to injection of the satellite into either

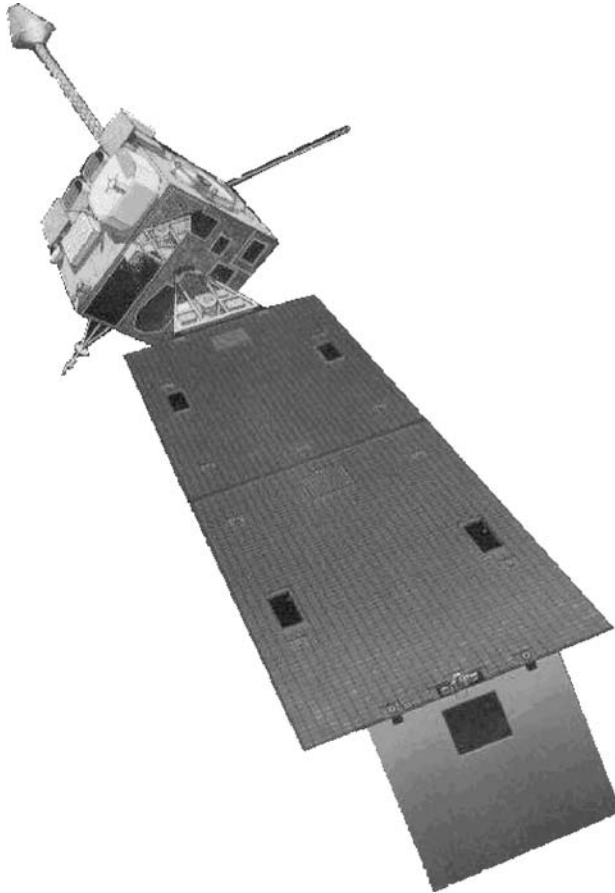


Figure 1. A geosynchronous operational environmental satellite (GOES). The function of this spacecraft is to make panoramic pictures of the world's weather patterns from an altitude of 22,500 miles.

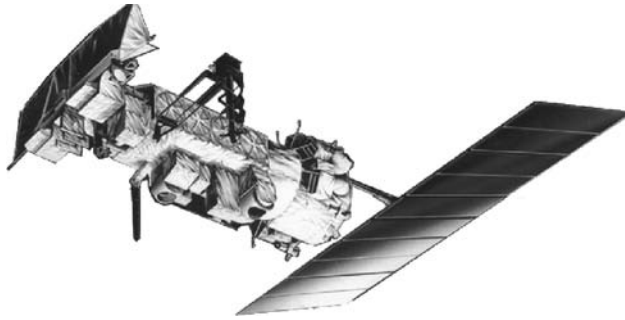


Figure 2. A polar-orbiting operational environmental satellite (POES). This spacecraft is a stable platform for high-resolution cameras and other sensors.

low-altitude polar or geostationary equatorial orbit and for the initial in-orbit satellite checkout and evaluation.

Early History of the Weather Satellite Program

Today's advanced technology and the images of clouds shown daily on television weather forecasts may make it difficult to remember the days when there were no weather satellites. Yet the need for weather observations from space was obvious. Oceans cover about 70% of Earth, and weather observations were sparse over these areas. Observations were limited even over the continents, except over North America and Europe.

The concept and design of weather satellites were developed during the 1950s. Even earlier, in the late 1940s, pictures of Earth taken from high-altitude V-2 rockets launched at White Sands, New Mexico, showed the meteorological potential through viewing cloud systems synoptically (1). These initial images demonstrated the need for satellites.

The earliest discussion of weather observations from "spaceships" was by Greenfield and Kellogg in a classified RAND report of 1951, later published in 1960 (2). The full potential of a satellite was recognized by Dr. Harry Wexler, chief scientist of the U.S. Weather Bureau (in the Department of Commerce, now part of NOAA). He presented his ideas at a symposium held at the Hayden Planetarium in New York City in 1954 (3). Wexler showed a simulated sketch of cloud cover over North America viewed from a satellite as if it were positioned 4000 miles over Texas. At the same symposium, Professor S. Fred Singer, then at the University of Maryland, presented his concept of a small, instrumented minimum orbiting unmanned satellite of Earth (MOUSE) to obtain a variety of atmospheric, geophysical, and astrophysical data, such as cosmic rays and ultraviolet radiation from the sun. Later, he developed these ideas in greater detail and published them in a review article in 1956 (4). Singer scanned Wexler's high-resolution cloud picture with an array of photocells on a simulated spinning satellite and showed a meteorologically adequate reproduction of cloud patterns (5). He also discussed measuring albedo (reflection of visible solar radiation) and

infrared (IR) emissions from the surface and atmosphere to determine long-term climate changes. The monograph also presented an analysis of an instrument to measure remotely the vertical distribution of ozone in the stratosphere (4).

By 1958, the United States began to launch a series of instrumented satellites through a newly established agency, the National Aeronautics and Space Administration (NASA) (6). It consolidated programs from many centers of the Department of Defense. In particular, it took over weather satellite design and testing from the U.S. Army Signal Corps. The program was headed by Dr. William Stroud and was transferred to the newly established Goddard Space Flight Center (GSFC) of NASA. The weather satellite, labeled TIROS, incorporated some of the design concepts of the MOUSE. It had a near-polar orbit to cover almost the entire globe; it was spin-stabilized, but it used television cameras rather than the scanning photodetectors of the simpler MOUSE.

In March 1958, the Chief of the USWB, Dr. F.W. Reichelderfer, established a special unit under Wexler, Director of Meteorological Research (7). He named Dr. Sigmund Fritz as the head of the Meteorological Satellite Section. Six months later, in September 1958, Wexler sent a memo to the USWB units noting the formation of NASA and stating that the Weather Bureau would be designated as its meteorological agent to provide meteorological instrumentation, data reduction, and analysis of observations. The memo also described in detail the types of data that were expected from the satellites and the research and development work needed. Almost 2 years before the launch of the first weather satellite, he identified the following types of data that would become available within the near future:

1. cloud cover
2. heat budget of the earth
3. nocturnal cloud cover using infrared
4. "temperature" as inferred from CO₂ emission in the infrared (IR)
5. ozone distribution
6. spectrum of solar radiation

The first TIROS launch demonstrated the operational value of cloud pictures, primarily for tracking hurricanes. It led directly to a Congressional authorization to set up an operational program in the U.S. Weather Bureau. NASA continued to hold the responsibility for research on satellites and instrumentation, and the Meteorological Satellite Section (MSS) of the USWB dealt with meteorological applications. As the agency responsible for operational applications, the USWB was to set the operational requirements that NASA had to satisfy in designing, launching, and operating the satellites.

The early TIROS carried two vidicon cameras—a wide-angle and a narrow-angle one. The image of the earth and clouds was formed on the vidicon tube (which is about 1 inch in diameter) and persisted for several seconds. During this period, the image was scanned and transmitted to the ground stations to receive and reconstruct the image. As soon as one image was transmitted, the next

sequence was started, and thus a large segment of the globe was covered. The satellite also carried a tape recorder to store the pictures when out of range of a ground data-acquisition station. The early satellites (TIROS-1 to -8 and TIROS-10) were shaped like a hatbox and were allowed to spin along their vertical axis to keep them stable. Because of the nature of the polar orbit and the way the cameras were mounted on the spinning satellite, only limited coverage of the globe was possible.

There were two primary ground stations to command and receive the satellite data. They were located at Ft. Monmouth, New Jersey, and Kaena Point, Hawaii. At the ground stations, pictures were displayed on kinescopes for immediate viewing and photographing. In the early stages of this program, NASA had the responsibility for planning; placing the satellite into orbit; and tracking, acquiring, processing, and analyzing the data. Assistance was provided by industry and other governmental agencies. Meteorologists in the USWB/MSS were responsible for analyzing and interpreting cloud-cover data.

When TIROS-1 was launched, it was assumed that the attitude of the satellite in space relative to the fixed stars would stay the same throughout its lifetime. Analysis of the pictures showed that this was not true and extensive corrections had to be made while analyzing the pictures. The causes for this deviation were (1) the interaction between the magnetic field of the satellite and that of Earth; and (2) the decrease of gravitational force with height, which causes an object in space to tend to align itself vertically. As soon as these factors were known, they were incorporated in future satellite data processing.

To observe both day and nighttime cloud cover, radiometers were installed on the next satellite, TIROS-2, and the follow-on satellites (TIROS-3, -4, -5 and -7). All of these radiometer measurements were stored in the tape recorder on board the satellite and were transmitted to the ground station after each orbit. Using the experience gained from these early measurements, various changes were made to the follow-on instruments, which will be discussed later.

The inclusion of infrared radiometers on TIROS added extra capabilities. For example, the measurements in the 8 to 12- μm channels provided an estimate of the cloud top temperatures, and thus one could infer the cloud top heights. One of the early studies by Rao and Winston (8) showed the cloud-top temperatures and cloud heights determined over the United States using the TIROS-2 radiation data. The researchers compared the results with synoptic and aircraft data and found good agreement.

Another interesting study (9) showed the temporal and spatial variations of the outgoing long-wave radiation derived from the TIROS-2 radiometer (7 to 33- μm channel) to estimate the total outgoing long-wave radiation at the top of the atmosphere. The observed latitudinal distribution agreed well with available theoretical studies.

These early satellites also mapped ice cover over the Gulf of St. Lawrence and the Great Lakes areas. These pictures were found very useful for navigation by observing the leads in the ice cover. Another important application was monitoring the extent of snow cover during winter and the melting of snow during spring. The snow cover information was very useful for hydrological forecasting (10,11).

Tiros-Nimbus Problems (PRE-1964)

Administrative History of NESDIS. *In 1960, the Meteorological Satellite Section, established in 1958 in the USWB, became the Meteorological Satellite Laboratory (MSL), and in 1962, the Meteorological Satellite Activities (MSA) with S. Fred Singer as the first Director and David S. Johnson as the Deputy Director. In the following year, it became the National Weather Satellite Center. When Singer departed in 1964, Johnson became Director. Upon the formation of the National Oceanic and Atmospheric Administration in 1972 and the accompanying reorganization, the organization was renamed the National Environmental Satellite Service (NESS). After Johnson's retirement in 1982, it became the National Environmental Data and Information Service (NESDIS). Dr. John McElroy was the Assistant Administrator until 1986. Dr. William Bishop served as Acting Assistant Administrator until Tom Pyke became the Assistant Administrator of NESDIS. He continued until 1992. After Pyke, Robert Winokur held the position until 1999, when Greg Withee became the Assistant Administrator of NESDIS.*

In June 1962, the Chief of the USWB appointed Dr. S. Fred Singer as Director of the Meteorological Satellite Activities. In a memo to his staff (Annex 5 of Ref. 7), he described the policies and the organization of the unit and noted in detail the "operational" responsibility for weather satellites bestowed on the USWB by the U.S. Congress in PL 87-332. Eighteen months after the launch of the first satellite TIROS-1, a fully operational entity was in existence.

As part of its responsibility, NASA developed the concept of the Nimbus research satellite. It was much larger, more complicated, and certainly more expensive than the simple TIROS. It was stabilized in three axes and was designed as a platform for a variety of instruments to view Earth by remote sensing. In addition to carrying out an atmospheric research program, Nimbus was also designed to provide a test bed for advanced instruments and systems, later to be transferred to the operational program (7).

Nimbus was primarily designed by Dr. Rudolph Stampf and his associates at NASA's Goddard Space Flight Center (GSFC) in 1959. The objectives of the design were (1) a near-polar orbit to permit observation of the entire Earth (from pole to pole), (2) Earth-stabilization so that the cameras and other sensors could always point toward Earth, (3) a retrograde (east to west) orbit inclined about 80° to the equator so that the satellite crosses the equator at local noon (northbound) and local midnight (southbound) in every orbit, (4) an altitude of 1000 km (600 nautical miles) to avoid Earth's radiation belt and to provide enough overlapping coverage at the equator, and (5) modular construction to allow easy exchange of sensors and communication modules. Originally, the plan was for GSFC to integrate and test the spacecraft. However, in 1961, several separate contracts were awarded for the various subsystems, and General Electric's Missile and Space Vehicle Department handled the final integration. Nimbus was a totally new system and was more complicated than the TIROS. The program was in trouble right from the beginning, and deadlines could not be met in completing the systems for integration and testing.

The second TIROS satellite was launched on 23 November 1960. Its ability to locate and track weather fronts, hurricanes and other tropical storms proved

to be a boon for meteorology. In addition to the USWB and its clients, the Department of Defense stated its interest in operational satellite data in no uncertain terms in Congressional hearings (12).

The evolution of the weather satellite program in the Department of Commerce is well described by P.K. Rao (7). He relates an interesting episode that took place in the early 1960s when the TIROS satellites were being considered for operational use. Some senior people in NASA wanted to consider it a research program and opposed immediate operational use of TIROS data. NASA refused the USWB use of the TIROS data for operational weather analysis for several weeks after the launch of TIROS. The two main groups that wanted the satellite data to be operational were the USWB and DOD Weather Services.

In October 1960, to settle some of the differences (research vs. operations) with regard to TIROS satellite operations, an interagency group consisting of NASA, USWB, DOD, and FAA established the Panel on Operational Meteorological Satellites (POMS) under the auspices of the National Coordinating Committee for Aviation Meteorology (NACCAM). The Chief of the USWB chaired NACCAM. The first chairman of POMS was Edgar Cortright of NASA; Dr. Morris Tepper of NASA served as secretary. This group developed the report "Plan for a National Operational Meteorological Satellite System (NOMSS)," completed in April 1961, and submitted to Congress (12).

POMS gave operational responsibility for weather satellites to the USWB, which then became the position of the Kennedy administration. Under this plan, Nimbus was to be the ultimate operational weather satellite but would also serve as a platform for developing instruments and technology. NOMSS was approved by President Kennedy as his fourth national goal; the first was to put a man on the Moon. On 25 May 1961, in President Kennedy's address to the Congress on "Urgent National Needs" (13), he requested funding for these activities (approximately \$53 million dollars) to be given to the USWB.

NOMMS said that the operational meteorological satellite system should

- satisfy the meteorological requirements of all users;
- phase into operation at the earliest date;
- capitalize on the continuing research and development program; and
- serve the United States first, but where possible, also serve international needs.

Already then, Stroud and his NASA colleagues foresaw problems as Nimbus was pushed into an operational role while still basically a research satellite. Because of the delay in Nimbus, new instruments could not be tested in time to incorporate them into operational weather satellites (7). NASA felt that Nimbus should be fully tested before declaring any instruments ready for operation.

Funds were appropriated for the USWB, which gave it fiscal control of the operational program, but NASA wanted to give responsibility to the USWB only when the system became operational. Nimbus, however, was a long way from becoming operational. As a result of the supplemental appropriation for the USWB of \$48 million for fiscal year 1962, the Nimbus Operational System (NOS)

agreement was signed; USWB gave up of most of its responsibility to NASA although it had the funds appropriated for it.

As Richard Chapman recounts in his history of the administrative, political, and technological problems of developing a U.S. weather satellite, the USWB was then in a weak position (12). It had no permanent director for its satellite activities and insufficient competence in the technical areas (outside of meteorology) to match those of the large NASA group at the GSFC. Change came in the spring of 1962, when the Commerce Department selected J. Herbert Hollomon for the position of assistant secretary for research and technology and in June 1962, when atmospheric and space physicist S. Fred Singer, then a professor at the University of Maryland, was appointed director of the weather bureau's satellite efforts (12).

On taking the position of director, Singer renamed the organization the National Weather Satellite Center (NWSC) and convened a group to correlate the requirements of operational users, including those of the DOD. The group was particularly concerned about the delays in the Nimbus program, that its capabilities had been scaled down, and that its lifetime and period of performance were short. Eventually, the delay lasted 2 years because of a variety of problems that GSFC as the integration contractor could not handle efficiently. At that time, Singer, as Director of the NWSC, decided to provide funds to build a few more TIROS satellites to maintain continuity of operations, rather than continue to fund NASA to develop Nimbus (12).

Gradually, a shift took place away from the NASA-USWB interagency agreement and NOMSS. Hollomon and Singer became particularly concerned about the accountability of USWB funds committed to NASA. What could the Department of Commerce and WB show the Congressional oversight committee in the way of accomplishments with its appropriated "no-year" funds? First, Singer decided that additional interim TIROS satellites were needed to fill the gap while waiting for Nimbus to become operational (12). Then, analyses by NWSC demonstrated that an advanced TIROS could handle some of the Nimbus functions, including APT, the all-important direct readout feature of cloud data that the DOD wanted. The main problem with Nimbus, aside from the delay, was its complexity and therefore the question of reliability for operational purposes. NASA itself only foresaw a lifetime of less than 6 months, whereas NWSC, concerned about cost, was looking for satellites that would operate for 3 to 5 years.

In the meantime, while the Nimbus cost had doubled, the USWB had reached the conclusion that an interim improved TIROS would be adequate and that something better than Nimbus was required for the long term. The NWSC studies showed that the cost of the operational program could be cut in half.

During the summer of 1963, a technical battle ensued between GSFC/NASA and the NWSC/ Department of Commerce, that covered many technical areas relevant to the operational weather satellite system, including optimum orbit, readout stations, instruments, and radiation protection.

The main NASA Command and Data Acquisition (CDA) station was near Fairbanks, Alaska. The turning point came when NWSC initiated its own technical studies and demonstrated that, in the proper orbit, the simpler satellite would require only one readout station (in Alaska) and that a Canadian station was not required. Ultimately, Singer did not accept the NASA analyses and

Secretary Hollomon formally withdrew support for the Canadian readout station, which was to be financed with money appropriated for the USWB. He argued that Nimbus had become too costly compared to an improved TIROS satellite. If NASA would not supply the spacecraft, then the DOD was willing to step in and manage the program (12).

This proved to be the final deciding factor. The TIROS operational satellite (TOS) was developed under NASA direction. It had many of the features of the Nimbus satellite and some of its instruments. However, it was simpler and cheaper. Instead of an active system of attitude control, it was a spin-stabilized wheel in a sun-synchronous, near-polar orbit. Most important, the budget of the operational program had been cut by nearly half, and Nimbus had lost its operational mission.

Nimbus 1 was finally launched from the Pacific Missile Range on 28 August 1964 on a Thor-Agena rocket but failed after 26 days. However, later and improved Nimbus satellites provided a test bed for instruments and technology that eventually found their way into the operational mission. The first Nimbus carried the advanced vidicon camera subsystem (AVCS) to provide global coverage. It also carried the automatic picture transmission (APT) subsystem to provide pictures of local cloud patterns directly to suitably equipped weather stations as the satellite passed over them. The AVCS and APT cameras provided images only during the daylight portion of the orbit. Nimbus also carried an improved version of a radiometer called the high-resolution infrared radiometer (HRIR) to observe in the 8- to 12- μm window region, and in the region of 3.4 to 4.2 μm (a clear window). The Nimbus also carried a radiometer, similar to the TIROS radiometer. This was called the medium-resolution infrared radiometer (MRIR) and again the window channel was changed to 10–11 μm so that the absorption due to ozone and water vapor could be eliminated (a much cleaner window region of the IR spectrum).

Between 1964 and 1979, NASA launched seven Nimbus satellites and tested prototype operational sensors for use in future NOAA polar-orbiting satellites (14).

Satellite Development—Early 1960 to Present

The early satellites TIROS-1 to -8 and TIROS-10 were shaped like a hatbox and were allowed to spin along their vertical axis to keep them stable. Because of the nature of the polar orbit (50° N–50° S) and the way the cameras were mounted on the spinning satellite, only limited coverage of the globe was possible. To obtain day and night coverage of the entire globe, radiometers were installed to observe in both the visible and infrared spectral regions. The rapid improvements in the spatial and spectral resolution of the radiometers made it possible to eliminate the onboard cameras. Both day- and nighttime images could be constructed from the radiometer measurements.

The first change in the weather satellite design occurred in 1965 in TIROS-9. The hat shape was changed to a cartwheel and the cameras were mounted along the radius 180° apart. The cameras could look straight down every time the satellite turned on its axis (at approximately 12 rpm). The spacecraft could view the entire Earth in a 24-hour period as it rolled along its orbit. The launch of the

TIROS-9 series increased the daily coverage capability of the satellite substantially. This was also the beginning of the TIROS operational system (TOS) satellites.

The First Tiros Operational System (TOS) Satellites. The first operational weather satellite system of the world started when the Environmental Sciences Services Administration (now NOAA) satellites, were launched ESSA-1, on 3 February 1966 and ESSA-2 on February 28, 1966. (The satellites reflect the name of the agency. The system consists of a pair of ESSA satellites in a sun-synchronous orbit; each is configured for a specific mission.) The advanced vidicon camera system (AVCS) obtained global imagery, which transmitted to the Command and Data Acquisition (CDA) stations at Wallops Island, Virginia, and at Fairbanks, Alaska. The CDA stations relayed the data to the National Environmental Satellite Service (NESS) in Suitland, Maryland, for processing and distribution to forecasting centers in the United States and other nations. The odd-numbered satellites (ESSA-1, -3, -5, -7, and -9) that had redundant AVCS systems were the global readout satellites. Even-numbered satellites (ESSA-2, -4, -6, and -8) were equipped with redundant automatic picture transmission (APT) cameras, and pictures from these cameras were directly transmitted to ground stations located around the world.

Improved Tiros Operational Satellite (ITOS) System (1970–1978). The new generation ITOS-1 satellite was launched on 23 January, 1970. The system carried both the AVCS and APT systems and a two-channel scanning radiometer (SR) that provided day and night coverage. The data from the SR were available by immediate transmission for local use and in a stored mode for later playback at the CDA station. The IR scanning radiometer made global observation of the atmosphere and surface areas available once every 12 hours from a single ITOS spacecraft. The second ITOS was launched on 11 December 1970, and it was named NOAA-1 (after the National Oceanic and Atmospheric Administration). The ITOS system evolved further from the development of ITOS-D satellites. These had a new redundant sensor complement to provide day and night imaging by the very-high-resolution radiometer (VHRR) and the medium-resolution scanning radiometer (SR). The VHRR and SR systems replaced the AVCS and APT cameras. The new ITOS system also carried the vertical temperature profile radiometer (VTPR) for obtaining the vertical temperature and moisture distribution in the atmosphere and a solar-proton monitor (SPM) for measuring solar protons and electron fluxes in the vicinity of the satellite. This second generation of satellite continued until 1978.

Third Generation of Operational Satellites: TIROS-N/NOAA-A to -D. These spacecraft covered a period from 1978 to 1981 and had a new and improved complement of systems. The advanced very-high-resolution radiometer (AVHRR) provided data for day and night imaging in the visible and infrared. It also provided observations to extract the sea surface temperature (SST), snow and ice distribution, and the Earth-atmosphere radiation budget. The TIROS operational vertical sounder (TOVS) provided vertical distribution of temperature and moisture in the atmosphere. The satellite also had a data collection system (DCS) to collect environmental data from stationary and moving platforms such as buoys, and remote hydrological stations. This satellite could broadcast data directly to local users and had a tape recorder to store data and transmit it to the

CDA stations at Wallops Island and Fairbanks. The TIROS-N series of satellites operated in polar orbit at an altitude of approximately 850 km.

The Advanced TIROS-N (ATN)/NOAA-E to -J System (1983–1994).

These spacecraft were modified to add some new and improved sensors: a search-and-rescue (SAR) system, Earth radiation budget experiment (ERBE), and a solar backscatter ultraviolet (SBUV) radiometer to measure stratospheric ozone distribution. The system also consists of two polar-orbiting satellites; operating as a pair, they provide environmental data for the entire globe, four times a day.

NOAA-K Series (1996 to Present). This is the latest in the Advanced TIROS-N series that started with NOAA-14. This satellite system carried a new version of the AVHRR-3 and is a six-channel instrument. It has an improved HIRS-3 instrument and an advanced microwave sounding unit-A (AMSU-A) to measure temperature and moisture from the surface to the upper stratosphere. The satellite also carried the British-built AMSU-B to measure the vertical distribution of water vapor in the atmosphere, a space-environment monitor (SEM) to measure the charged particles entering Earth's atmosphere, SAR instruments, and data collection instruments.

Geostationary Environmental Satellites—History. The NOAA GOES program was a direct outgrowth of NASA's Applications Technology Satellite (ATS) program. It was initiated in 1966 to demonstrate communications technology by using a satellite in a geostationary orbit. The major objective of the early ATS satellites was to test whether gravity would anchor the satellite in a 24-hour synchronous orbit (22,300 miles above Earth's surface) over the equator, allowing it to orbit at the same rate as Earth turns, thus seeming to remain stationary. The excess capacity of the spacecraft allowed including meteorological sensors for experimental observations of Earth from a geosynchronous altitude. A spin-scan camera developed by Professor V. Suomi of the University of Wisconsin provided continuous images of the sunlit Earth disk every half hour. The nearly continuous imagery proved the ability of the spacecraft to monitor the evolution of weather systems in real time, particularly severe weather. Another satellite in this series, ATS-3, was launched in 1967 to cover the Western Hemisphere. The next in this series, ATS-6, was totally different; it was a three-axis-stabilized spacecraft intended primarily for communication experiments from geostationary orbit. It carried a meteorological sensor, the geostationary very-high-resolution radiometer (GVHRR), a two-channel radiometer scanning in the visible range (0.55–0.75 micron) and the IR (10.5–12.5 micron). Because of the IR, it was possible to image during the day and night.

Synchronous Meteorological Satellites. NASA launched two synchronous meteorological satellites (SMS), SM-1 in May 1974 and SMS-2 in February 1975, to demonstrate their use for weather forecasting. After the successful launch of these satellites, NASA turned the program over to NOAA for operation, and they were renamed geostationary operational environmental satellites (GOES).

Geostationary Operational Environmental Satellites (GOES). GOES satellites are a mainstay of weather forecasting in the United States. They are the backbone of short-term forecasting or "now casting." The real-time weather data gathered by GOES satellites, combined with data from Doppler radars and automated surface observing systems, greatly aid weather forecasters in

providing warnings of thunderstorms, winter storms, flash floods, hurricanes, and other severe weather. These warnings help to save lives and preserve property.

The United States operates two meteorological satellites in geostationary orbit, one over the East Coast and one over the West Coast, that give overlapping coverage of the United States. Currently, GOES-8 and GOES-10 are in operation. The GOES satellites are a critical component of the ongoing National Weather Service modernization program, aiding forecasters in providing more precise and timely forecasts. The next GOES-11 satellite (GOES-L) launched in 2000 is stored in orbit, and GOES-12 launched in 2001 is also stored in orbit. They will be activated when one of the current GOES satellites fails. They are the first of the NOAA satellites equipped with a solar X-ray imager (SXI), an instrument that can detect solar storms.

Defense Meteorological Satellite Program (DMSP). Since the mid-1960s, the U.S. Air Force has operated polar-orbiting meteorological satellites. The observations have emphasized high-resolution and low-light imaging rather than atmosphere profiling. The DMSP system maintained two satellites at any given time that crossed the equator in midmorning and late evening. Some of the later satellites carried microwave imaging and sounding units; some improved versions of these instruments were incorporated in the NOAA polar operational satellites. So far, about 20 DMSP satellites have been flown, primarily for use by the Defense Department.

International Program Cooperation. In the 1980s, NOAA had to balance the high cost of space systems and the growing need to provide a complete and accurate description of the atmosphere at regular intervals as inputs to numerical weather prediction and climate monitoring support systems. This led NOAA to enter into discussions and agreements at the international level with the European Organisation for the Exploitation of Meteorological Satellites (EU-METSAT). The goal of this cooperation is to provide continuity of measurements from polar orbits, cost sharing, and improved forecast and monitoring capabilities by introducing new technologies.

Several countries recognized the advantages of monitoring Earth and its environment from space and have launched both polar-orbiting and geosynchronous satellites. The operators include the European Space Agency (a consortium of several European countries), Japan, Russia, India, and China. Under an agreement reached by the satellite operators under the auspices of the World Meteorological Organization (a U.N. agency in Geneva, Switzerland), the locations of geosynchronous satellites were distributed to provide global coverage. Thus, the two U.S. satellites, a European satellite, a Japanese satellite, and a satellite from India provide complete coverage continuously from about 60° N to 60° S. Satellite data sets and products were coordinated to ensure maximum compatibility among the operators of the geosynchronous satellites.

Russia (the former Soviet Union) has launched a series of polar-orbiting satellites in the Meteor series to obtain cloud cover, snow and ice extent, and the Earth radiation budget, and also a few sounding instruments to measure the vertical distribution of temperature and moisture for weather prediction models.

The European Space Agency is planning to launch a new generation of polar-orbiting weather satellites that will be discussed under the future satellite program.

Remote Sensing

Satellites by their nature have to rely on remote sensing. The fundamentals are elaborated in a number of books (15–21). We will discuss the subject in four topics:

1. reflection and scattering of solar radiation
2. thermal emission from the surface and from the atmosphere
3. active probing using radar and lidar
4. other applications

Reflection and Scattering of Incident Solar Radiation. This can take place at wavelengths ranging from the near-infrared (IR) around 3 microns down to the near-ultraviolet of about 0.28 microns. Beyond 2 microns, there are important IR absorption bands, principally from water vapor, carbon dioxide, and from other minor atmospheric constituents (Fig. 3).

The most important reflection comes from clouds in the atmosphere and snow and ice on the surface. The reflection coefficient, the so-called albedo, approaches 100% in many cases. On the other hand, the albedo of the ocean is less than 10% except for areas of “sun glint” where the reflection is specular. The patterns of reflection and their changes with time lend themselves to meteorological analysis, as discussed in the next section.

In addition to meteorological applications for weather prediction, observations of so-called aerosols, including dust, smoke, and even locust swarms, yield important information for a variety of purposes. In addition, spectral data at different wavelengths (color) can give information about agricultural crops, the health of forest systems, ocean productivity, and even about the mineral content

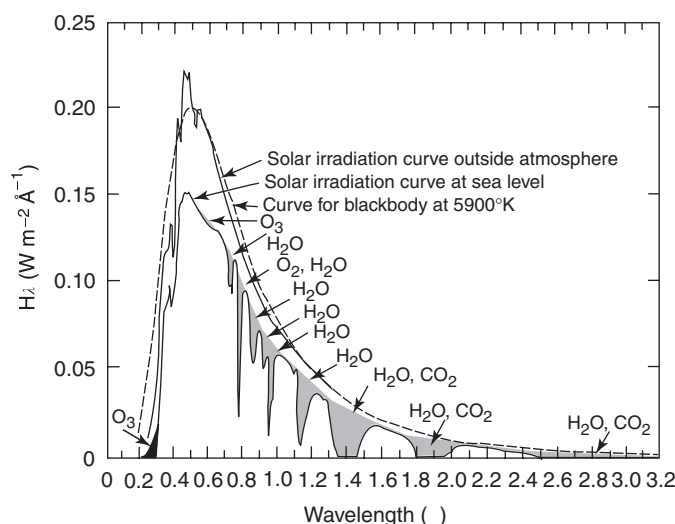


Figure 3. Spectral curves of solar radiation incident at the top of the atmosphere and at the surface of Earth for the Sun at zenith on a clear, cloud-free day (courtesy of Air Force Cambridge Research Laboratories, 1965).

of the soil (10,11). As discussed in specialized texts, measurement of the polarization of reflected and scattered light yields additional information about the optical and physical properties of the scattering medium.

Backscattered radiation in the near UV region between 0.28 and 0.32 microns can be used to measure the amount and vertical distribution of stratospheric ozone. This technique was analyzed before satellites were launched (4) and was first tested on Nimbus. It has now become operational and is used to monitor ozone worldwide, including tracking the Antarctic ozone hole. Sulfur dioxide gas emitted by volcanoes into the stratosphere also absorbs strongly in the near ultraviolet; but by choosing the right wavelengths and adequate spectral resolution, one can distinguish between ozone and sulfur dioxide.

Emission from the Atmosphere and Surface. Corresponding to the lower temperature of the surface and atmosphere, compared to the Sun, the emission wavelength range covers from about 4 microns into the far infrared beyond 20 microns and also the microwave region. The important feature here is the so-called "window" of the atmosphere. In the absence of clouds, IR emission from the surface in the region between 8 to 12 microns can penetrate the atmosphere and escape into space where it is observed by a satellite. (Fig. 4). Outside the window region, the atmosphere is opaque in the infrared because of strong absorption bands, principally from water vapor and carbon dioxide. (There is also absorption by stratospheric ozone at 9.6 microns in the window region.) Water droplets absorb and emit strongly in the infrared so that IR detectors from the satellite view only the tops of clouds. Because temperature decreases sharply in the troposphere, the IR emitted from cloud tops can be used to obtain an estimate of the pressure and altitude of the cloud tops.

The great advantage of infrared emissions is that they can be measured at night and are easier at that time. During the day, there is some interference from solar radiation in the infrared. In the absence of clouds, under clear sky conditions, one can measure the surface temperatures of the ground and soil and get some determination of soil moisture from the diurnal variation. Information from the surface of the ocean provides a measure of temperature; however, it relates to the skin of the sea surface and is heavily influenced by the sea state, foam, wave action, and therefore surface winds. In addition, one has to be careful in comparing IR measurements of SST with measurements from ships because they refer to different layers of the upper ocean.

When the atmosphere is clear, and there are no clouds, infrared emissions from the 15-micron band of carbon dioxide can be used to derive the vertical distribution of temperature in the atmosphere. This method was first suggested by Lewis Kaplan and was worked out in greater detail by David Wark and others at the NWSC and by Rudolph Hanel and Wm Bandeen at NASA. Mathematically, the problem is that of determining the kernel of an integral, when the integrals are observed at several wavelengths around 15 microns. This deconvolution problem also enters into the determination of atmospheric temperature distribution from microwave measurements. (The latter have an advantage over IR measurements because they can be carried out in the presence of clouds.) A similar problem occurs in measuring the vertical distribution of ozone (4); it is somewhat simpler because in the UV, only scattering and absorption are important but not emission.

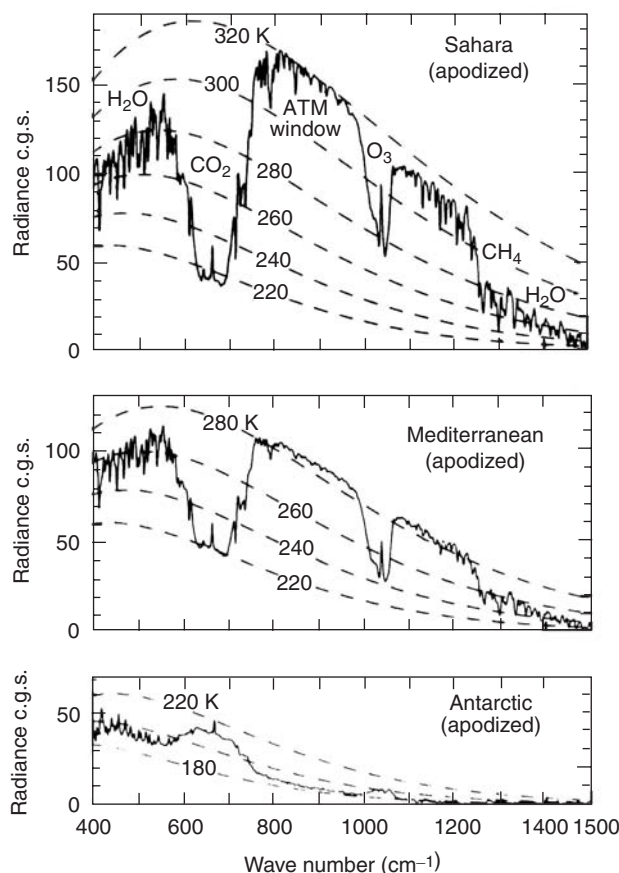


Figure 4. Spectra of the radiation emitted by the Earth-atmosphere system, as seen by the Nimbus satellite above three different points on Earth. Radiance units (cgs) are $\text{erg} \cdot \text{s}^{-1} \cdot \text{cm}^{-2} \cdot \text{sr}^{-1} / \text{cm}^{-1}$ (from Conrath, B.J., et al. Vertical sounding of the atmosphere with the Nimbus IV infrared experiment. *Proc. 21st Astronaut. Congr.*, 1971, North Holland, p. 1010, Fig. 2).

Infrared radiation from the surface can provide interesting information on ground cover vegetation and can measure desertification. The Hyperion instrument on the NASA Earth Observing-I satellite, launched in November 2000, collects reflected radiance in 220 spectral bands covering the range from 0.4–2.5 microns. Preliminary research results are reported in Ref. 22; it is not clear yet whether this will lead to operational use of the technique.

Measuring the emission of microwaves from the surface and from the atmosphere has proven to be immensely useful. An early experiment in 1964 measured emission from the sea surface and demonstrated that the energy emitted, hence the emissivity, depended primarily on the sea state (23). Microwaves can measure such quantities as snow cover, precipitation, and thunderstorms (24,25). Thermal microwave emission from water vapor can be used to measure atmospheric humidity, an important quantity for studying global

climate change. Microwave emission from molecular oxygen can be used to measure the bulk temperature of the troposphere and the lower stratosphere, as demonstrated in careful analyses by Christy and Spencer (26). As discussed later, satellite microwave data indicate that the atmosphere has not warmed perceptibly since global temperature observations began in 1979.

Active Probing with Lasers and Radar. Radar reflection from the sea surface can determine the sea state and thereby surface winds. This has become an operational application because wind field patterns play an important role in weather prediction. Radar reflection from the land surface can produce accurate determinations of topography, including even snow depth. Finally, radar can be used to measure precipitation, especially useful over the oceans where other observations are not available. Satellite-borne synthetic aperture radar (SAR) has been used in many applications (see the Applications Section below). A recently reported use is detecting of urban sewage and storm water runoff from urban areas that cause marine pollution (27).

Radar scatterometers have provided nearly continuous coverage of Earth since 1991 on satellites of NASA, European Space Agency (ESA), and Japan. Frequencies have ranged from 5.3 to 14.6 GHz, and both vertical and horizontal polarization were measured. Applications include Greenland and Antarctic ice sheets, sea ice, soil moisture, and vegetative coverage (28,29).

Laser measurements are still in their infancy. Ideally, one would like to measure the Doppler shift of reflections from particles in the atmosphere and thereby deduce horizontal wind velocities.

Another application, still on the drawing board, uses the absorption in the D-band of oxygen molecules around 0.76 microns to measure surface pressure in the absence of clouds (30). When clouds are present, the technique measures the pressure and altitude of cloud tops. Combined with the temperature of the cloud tops, this amounts to vertical probing of the atmosphere. Systems analysis shows that internal noise is negligible and that background is not serious, even in daylight. Compared with the corresponding passive method using the Sun as a source, the laser method can be used at night, can discriminate cloud versus surface reflections, and may be able to determine altitude, pressure, and (by IR flux measurement) the temperature at selected points in the atmosphere. If successful, the method will have important applications to cloud studies and to oceanography.

Other Applications of Remote Sensing. More refined measurements, based on wavelength dependence, polarization measurements, and time variability, can be used to deduce such important quantities as soil moisture, ocean productivity, and the properties and nature of aerosols, such as size distribution and their optical parameters (17,18,21). Besides weather satellites, other kinds of satellites can also supply data important for meteorology. For example, the Global Positioning System can be used to measure upper tropospheric water vapor by studying the time delays at the two different radio frequencies used by GPS. The TOPEX satellite can be used to measure the height of the sea surface, from which one can derive the location and strength of ocean currents. Finally, astrophysical satellites can measure solar radiation, its variability in the visible and ultraviolet, and the nature and flux of solar particulate radiation impacting on Earth's atmosphere.

Applications

The earliest meteorological results from TIROS were published in three articles soon after launch (31–33). By the end of April 1960 (almost 30 days after launch), a number of case studies of meteorological phenomena observed by TIROS were begun by a team of meteorologists in the Meteorological Satellite Section of the USWB. These studies included several large-scale cyclonic vortices over the United States, the North Atlantic, and the North Pacific; cloudiness in the tropical regions of the South Pacific; cellular arrangements of cumulus clouds over the Atlantic and Pacific Oceans in temperate latitudes; cloud streets in the Caribbean, cloudiness associated with severe thunderstorms and tornadoes; ice in the Gulf of St. Lawrence; orographic clouds in various parts of the world; snow cover in mountain regions; and sun glitter on the ocean surface. Several technical reports were published by USWB and NASA in the following months (7).

The TIROS pictures received at the ground stations were recorded on 35-mm film by a kinescope camera, either during a satellite readout or by playing back the data recorded on magnetic tape. The film was processed immediately to make transparencies for projection and for prints. Geographic reference grids were overlaid on these pictures to determine the location where the pictures were taken. These overlay grids were generated by taking into account the position of the satellite, the time when the picture was taken, the direction in which the camera was pointing, and other parameters connected with the satellite spin axis. Once the grids were overlaid on the pictures, cloud analyses (called neph-analyses) could be performed, showing cloud types and extent of coverage. These maps were sent by facsimile to weather stations around the globe for immediate use. Within 48 hours of the TIROS-1 launch, such pictures and nephanalyses were made available to USWB meteorologists, the U.S. Air Weather Service, and the U.S. Naval Weather Service. The information was limited to areas where sunlight was available.

The introduction of infrared sensors in radiometers on TIROS-2 and the follow-on satellites, starting from November 1960, enabled continuous day and nighttime coverage. Three types of instruments were used: (1) a five-channel scanning radiometer, (2) a two-channel medium-resolution radiometer, and (3) a two-channel omnidirectional radiometer. Of the three instruments, the five-channel one was the most important. It consisted of the following:

1. 5.9–6.7 μm to measure atmospheric water vapor (based on strong absorption in this spectral region);
2. 8 to 12- μm atmospheric window to measure surface temperatures under cloud-free conditions and cloud temperatures and heights and to obtain nighttime images;
3. 0.2–5 μm to measure the albedo of the earth, clouds, etc. (reflectance);
4. 7.0–30 μm to measure the total outgoing long-wave radiation at the top of the atmosphere;
5. 0.5–0.7 μm in the visible part of the spectrum to measure the reflected radiation and create images during the daytime.

The sensors were mounted in the satellite at an angle of 45° to the spin axis of the satellite. The scan was produced by the spin of the satellite. The movement of the satellite provided coverage along the track. Because of the spin of the satellite, the sensors looked alternatively at space and at Earth. The field of view of the radiometer was approximately 30 miles.

All of these radiometer measurements were stored in the tape recorder on board the satellite and transmitted to the ground station after each orbit. The data were processed after reception at the ground station. Based on the experience gained from these early measurements, various changes were made in the follow-on instruments, which will be discussed later.

The inclusion of infrared radiometers on TIROS added extra capabilities. For example, the measurements in the 8- to $12\text{-}\mu\text{m}$ channels provided an estimate of cloud top temperatures, and thus one could infer the cloud top heights. One of the early studies by Rao and Winston (8) showed the cloud top temperatures and cloud heights determined over the United States using the TIROS-2 radiation data. The researchers compared the results with synoptic and aircraft data and found good agreement. It was also pointed out in this paper that the 8- to $12\text{-}\mu\text{m}$ window region is not clean and there is a considerable amount of absorption due to water vapor and ozone; thus appropriate corrections needed to be applied.

Satellite pictures were also used to observe and map the ice cover over the Gulf of St. Lawrence and the Great Lakes. These pictures were useful for navigation by showing the leads in the ice cover. Another important application was to monitor the extent of snow cover during winter and the melting of snow during spring. The snow cover information was useful for hydrological forecasting.

In the early days of the satellite program, it was realized that in addition to the cloud information, the temperature and moisture distribution with height in the atmosphere is essential for forecasting. Numerical modeling of the atmosphere was gaining momentum at this time, and observations over the vast oceanic regions and other data-void regions of the world were needed to fill the gaps in these models. A prototype instrument called the satellite infrared spectrometer (SIRS) was under development within the Meteorological Satellite Laboratory. The first sounder was on the Nimbus satellite launched in 1969. Observations from the SIRS instrument were used to derive the vertical distribution of temperature (and, to a limited extent, moisture) over the oceanic regions. The satellite-derived soundings were useful for numerical weather forecast models, particularly over the oceanic regions, where radiosonde data are not available. NASA was also developing another sounder at the same time, called the vertical temperature profiling radiometer (VTPR). Because of the high spectral resolution of the VTPR, it was superior to the SIRS. Immediate plans were made to improve the VTPR instrument to operational status and move it to the NOAA polar satellites. The first operational sounder was launched on NOAA-2 in 1972. The sounders continued on all polar-orbiting satellites from that period.

In addition to the visible and infrared portions of the spectrum, satellites also started to exploit the ultraviolet, using the backscatter technique to measure the amounts and vertical distribution of ozone. Such instruments as TOMS and

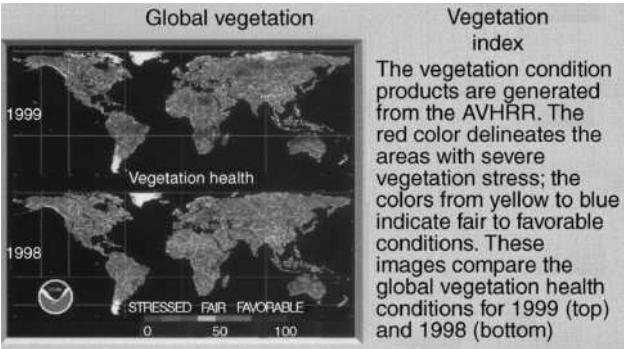


Figure 5. Vegetation Index: Vegetation condition products are generated from the AVHRR. The red color delineates the areas that have severe vegetative stress; the colors from yellow to blue indicate fair to favorable conditions. These images compare the global vegetative health conditions for 1999 (top) and 1998 (bottom). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

SBUV were first flown on Nimbus. They were then transferred to the operational weather satellites.

Satellites have also exploited another region of the electromagnetic spectrum, measuring the passive (thermal) emission of microwaves from the surface and from atmospheric components. An early experiment, conducted from a blimp over Biscayne Bay, Florida, demonstrated the high emissivity from foam produced by waves (and thus the sea state). Nowadays, active microwave (radar) determines the sea state by scattering.

A comprehensive account of weather satellite applications is given in various references. Here we will provide two examples to demonstrate the wide range of results obtainable by remote sensing from satellites. Both of these images were made using the advanced very high resolution radiometer on POES spacecraft. Figure 5 is a composite picture of Earth showing the state of worldwide vegetation. Figure 6 is a picture of Hurricane Floyd. This is an example of

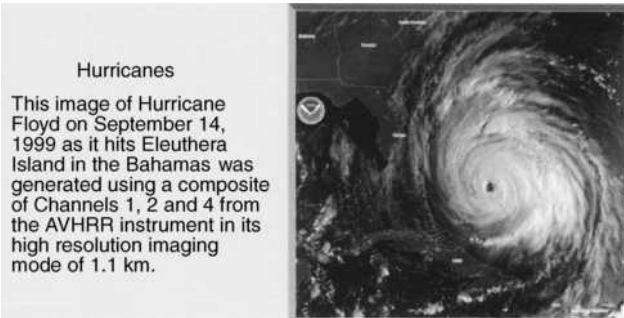


Figure 6. Hurricanes: This image of Hurricane Floyd on 14 September 1999 as it hits Eleuthera Island in the Bahamas was generated using a composite of channels 1, 2, and 4 from the AVHRR instrument in its high-resolution imaging mode of 1.1 km. This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

the high-resolution pictures of various weather patterns that can be made by using satellite-borne instruments.

Climate Studies

Weather satellite instruments were not designed to determine long-term trends of atmospheric parameters. Changes over decadal timescales generally require using instruments from different satellites and careful calibration. Yet, with great care it has been possible to use much of the data for such purposes.

Four kinds of measurements are possible; the first two refer to the atmospheric energy balance:

1. Albedo measurements over long time spans can, in principle, determine any changes in the energy input to Earth. Calibration is quite difficult here, but attempts have been made to establish long-term trends in cloudiness and surface albedo through changes in human land use (34).
2. The converse measurement measures the outgoing long-wave radiation (OLR) from Earth, including the IR emitted from the surface through the atmospheric window and the IR emitted from the atmosphere, mainly from upper tropospheric water vapor, carbon dioxide, and clouds (35,36).
3. Rather than concentrate on the energy balance, a complementary technique monitors long-term changes in global atmospheric temperature. For this purpose, the microwave sounding unit (MSU) carried by weather satellites has turned out to be the principal instrument (26). There is no perceptible trend in global mean temperature since observations started in 1979. Surface observations from weather stations do show a warming during the same period (37); this discrepancy has not yet been resolved (38). However, independent data from radiosondes in weather balloons confirm the satellite data that there has been no atmospheric warming. Proxy data of surface temperature from tree rings and ice cores also show no surface warming trend since about 1940. Thus the preponderance of data indicates little if any warming in the last 60 years.
4. Finally, thanks to satellites, it may be possible to measure long-term changes in global precipitation, particularly over the oceans where there are no good data. We do have long-term measurements of severe storms and tropical cyclones; land observations and also satellites have so far shown no clearly established trends (37). As reported, the intensity and frequency of North Atlantic hurricanes has decreased. There has been no report of discernible trends for severe storms, El Niños, and similar meteorological phenomena.

Continuity and Future of Operational Satellite Programs

Figure 7 shows the planned launch of polar-orbiting and geostationary operational satellites. It should be emphasized that the actual schedules will depend primarily on the need for replacement satellites, the availability of spacecraft,

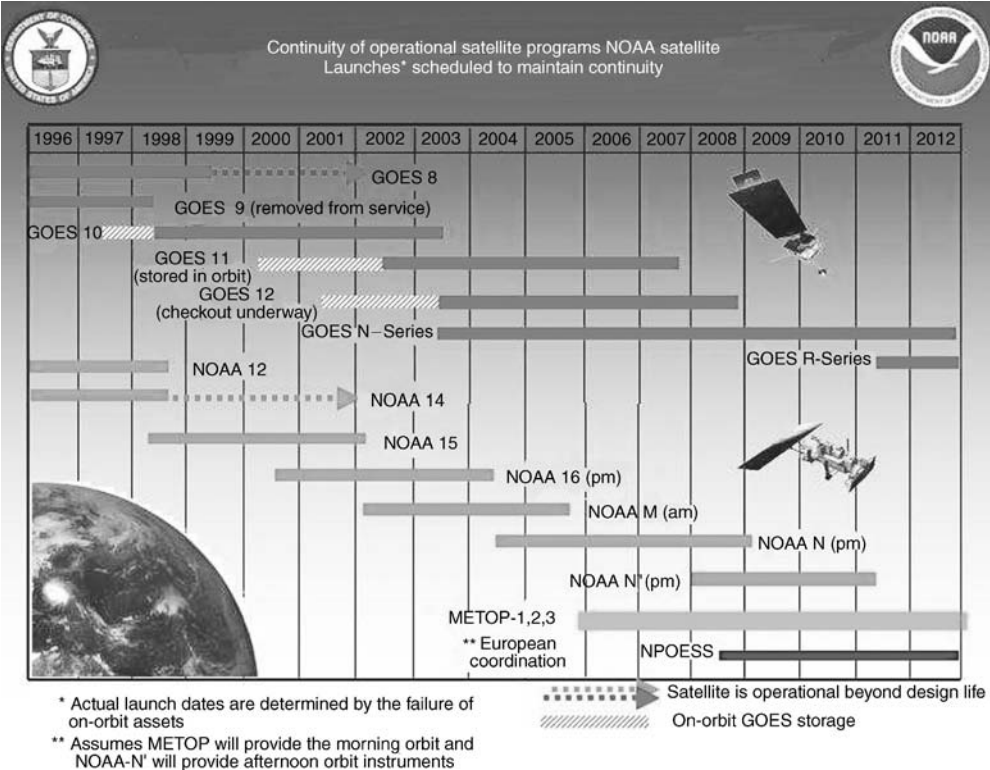


Figure 7. Continuity of Operational Satellite Programs. NOAA satellite launches scheduled to maintain continuity (courtesy of NOAA). This figure is available in full color at <http://www.mrw.interscience.wiley.com/esst>.

and a launch vehicle. It is also assumed that the new generation of satellites will have a 5-year life expectancy and that all launches are successful. It should be pointed out that the satellites designated by letter are designated by a number after a successful launch (e.g., GOES-N to GOES-13; NOAA-M to NOAA 17).

The U.S. government has traditionally maintained two operational weather satellite systems; each has a 30-plus year history of successful service: NOAA's polar-orbiting operational environmental satellite (POES) and DOD's Defense Meteorological Satellite Program (DMSP). Recent changes in world political events and declining agency budgets prompted a reexamination of combining the two systems.

In May 1994, President Clinton signed a presidential directive to merge the civilian POES system and the DMSP into a single system to reduce costs, while continuing to satisfy the U.S. operational requirements. On 3 October 1994, NOAA, DOD, and NASA created an Integrated Program Office (IPO) to develop, manage, acquire, and operate the NPOESS system. The Integrated Program Office concept provides each of the participating agencies with lead responsibility for one of three primary functional areas. NOAA has overall responsibility for the merged system and is also responsible for satellite operations. NOAA is also the primary interface with the international response and civil user communities.

DOD is responsible for supporting the IPO for major systems acquisitions, including launch support. NASA has primary responsibility for facilitating the development and incorporation of new cost-effective technologies into the merged system. Although each agency provides certain key personnel in their lead roles, each functional division is staffed by triagency work teams to maintain the integrated approach.

As an early step in the merging process and the first tangible result of the NPOESS effort, Satellite Control Authority for the existing DMSP satellites was transferred in May 1998 from the U.S. Air Force Space Command to the NPOESS Integrated Program Office. The command, control, and communications functions for the DMSP satellites have been combined with the control for NOAA's POES satellites at NOAA's Satellite Operations Control Center (SOCC) in Suitland, Maryland. The DMSP satellites are being "flown" by civilian personnel at the SOCC. This is the first time in the 30-plus-year history of this DOD program that DMSP satellites have not been flown by Air Force personnel. A backup satellite operations center manned by USAF crews is also contemplated.

On 13 December 1999, a new Department of Defense meteorological satellite was launched by the U.S. Air Force and is being operated by NOAA. The satellite is the next in a series of the Defense Meteorological Satellite Program. This is the first DMSP whose postlaunch checkout was conducted from NOAA's Satellite Operations Control Center in Suitland, Maryland.

The NPOESS Preparatory Project (NPP) is a proposed joint mission to extend key measurements in support of long-term monitoring of climate trends and of global biological productivity. It extends the measurement series being initiated with EOS Terra (MODIS) and EOS PM (AIRS, AMSU, HSB) by providing a bridge between NASA's EOS missions and the NPOESS system. The NPP mission will provide operational agencies with early access to the next generation of operational sensors, thereby greatly reducing the risks incurred during the transition. This will permit testing of advanced ground operations facilities and validation of sensors and algorithms while the current operational systems are still in place. This new system will provide nearly an order of magnitude more data than the current operational system. Launch is planned for late 2005. As a result of 5-year design lifetime, NPP will provide data past the planned lifetime of EOS Terra and EOS PM and into the expected launch of the first NPOESS satellite. The proposed NPP mission is currently in the formulation phase.

The first merged NPOESS satellite is expected to be available for launch in the latter half of the decade, approximately 2008, depending on when the remaining POES and DMSP program satellite assets are exhausted. NPOESS will provide significantly improved operational capabilities and benefits to satisfy the nation's critical civil and national security requirements for space-based, remotely sensed environmental data. NPOESS will deliver higher resolution and more accurate atmospheric and oceanographic data to support improved accuracy in short-term weather forecasts and warnings and severe storm warnings, as well as to serve the data continuity requirements of the climate community for improved climate prediction and assessment. NPOESS will also provide improved measurements and information about the space environment necessary to ensure reliable operations of space-based and ground-based systems

and will continue to provide surface data collection and search and rescue capabilities.

NPOESS Instruments. The NPOESS satellites will carry six instruments designed to revolutionize weather forecasting and climate research:

1. Visible/infrared imager/radiometer suite (VIIRS). This sensor is arguably the most important in the entire NPOESS blueprint. It will spot smoke from forest fires, ash from volcanoes, and the rain bands of hurricanes. Images from VIIRS will show up on the nightly news and the Weather Channel. VIIRS will replace the advanced very-high-resolution radiometer (AVHRR) on the existing POES satellites. VIIRS will divide light into 22 channels, compared to six for AVHRR.
2. Conical microwave imager/sounder (CMIS). The advantage of CMIS is that it can see through clouds. It will detect microwave radiation emitted from the surface of the ocean and from the atmosphere. It will replace the special sensor microwave imager on DMSP and the microwave imager on NASA's TRMM satellite.
3. Crosstrack infrared sounder (CrIS). CrIS is the primary instrument for measuring the temperature, moisture, and pressure of the atmosphere. It will replace the high-resolution infrared sensor on POES.
4. GPS occultation sensor (GPSOS). GPSOS will measure the atmospheric refraction of radio signals from the GPS satellite constellation and from the Russian Global Navigation Satellite System.
5. Ozone mapping and profiler suite (OMPS). OMPS measures the vertical and horizontal distribution of ozone. It will aid studies of the seasonal ozone hole over Antarctica and the thinning ozone layer over the Arctic.
6. Space environment sensor suite (SESS). SESS measures particles in the upper reaches of the atmosphere that can damage satellites. It will measure neutral and charged particles, electrons, and magnetic fields, and the optical signatures of the aurora phenomenon.

The European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) has also joined the consortium and will provide the first European meteorological operational satellite in polar orbit, called METOP, in 2004. It is planned to carry many instruments provided by the United States and will provide early morning and late evening coverage. Initially three satellites are planned in this series (39,40).

ACRONYMS

AIRS. Advanced Infrared Sounder

AMSU. Advanced Microwave Sounding Unit

APT. Automatic Picture Transmission

ATS. Applications Technology Satellite

AVCS. Advanced Vidicon Camera System

ATN. Advanced TIROS-N

AVHRR. Advanced Very High Resolution Radiometer
CDA. Command and Data Acquisition
CMIS. Conical Microwave Imager/Sounder
CrIS. Crosstrack Infrared Sounder
DOD. Department of Defense
DMSP. Defense Meteorological Satellite Program
EOS. Earth Observing System
ERBE. Earth Radiation Budget Experiment
ESA. European Space Agency
ESSA. Environmental Science Services Administration
EUMETSAT. European Organisation for the Exploitation of Meteorological Satellites
FAA. Federal Aviation Administration
GOES. Geostationary Operational Environmental Satellite
GPS. Global Positioning System
GPOS. GPS Occultation Sensor
GSFC. Goddard Space Flight Center
GVHRR. Geostationary Very High Resolution Radiometer
HRIR. High Resolution Infrared
IPO. Integrated Program Office
IR. Infrared
ITOS. Improved TIROS Operational System
MOUSE. Minimum Orbiting Unmanned Satellite of Earth
MRIR. Medium Resolution Infrared
MSA. Meteorological Satellite Activities
MSL. Meteorological Satellite Laboratory
MSS. Meteorological Satellite Section
MSU. Microwave Sounding Unit
NACCAM. National Coordinating Committee for Aviation Meteorology
NASA. National Aeronautics and Space Administration
NESDIS. National Environmental, Satellite, Data, and Information Service
NESS. National Environmental Satellite Service
NOAA. National Oceanic and Atmospheric Administration
NOMMS. National Operational Meteorological Satellite Systems
NOS. Nimbus Operational System
NPOESS. National Polar-Orbiting Operational Environmental Satellite System
NPP. NPOESS Preparatory Project
NWSC. National Weather Satellite Center
OLR. Outgoing Long-wave Radiation
OMPS. Ozone Mapping and Profiler Suite
POES. Polar-Orbiting Environmental Satellite
POMS. Panel on Operational Meteorological Satellites
SAR. Search and Rescue
SBUV. Solar Backscattered Ultraviolet

SESS. Space Environment Sensor Suite
SEM. Space Environment Monitor
SIRS. Satellite Infrared Spectrometer
SMS. Synchronous Meteorological Satellite
SOCC. Satellite Operations Control Center
SR. Scanning Radiometer
SST. Sea Surface Temperature
TIROS. Television Infrared Observation Satellite
TOMS. Total Ozone Mapping Spectrometer
TOS. TIROS Operational Satellite
TOVS. TIROS Operational Vertical Sounder
TRMM. Tropical Rainfall Measurement Mission
USWB. United States Weather Bureau
UV. Ultraviolet
VIIRS. Visible Infrared Imager Radiometer Suite
VTPR. Vertical Temperature Profiling Radiometer

BIBLIOGRAPHY

1. Holliday, C., In *The First Forty Years: a Pictorial Account of The Johns Hopkins University Applied Physics Laboratory Since Its Founding in 1942*. The JHU Applied Physics Laboratory, Laurel, MD., 1983. Chapter 2, pp. 13–18 (1946).
2. Greenfield, S.M., and W.W. Kellogg. Inquiry into the Feasibility of Weather Reconnaissance from a Satellite Vehicle. RAND Corporation Report, No. R-365, (unclassified edition of RAND Report No. R-218, issued 1951) 1960.
3. Wexler, H. Observing the weather from a satellite vehicle. *J. Br. Interplanetary Soc.* 14: 269–276 (1956).
4. Singer, S.F. Geophysical research with artificial earth satellites. In H.E. Landsberg (ed.), *Advances in Geophysics*, Academic Press, New York, 1956, Vol. 3.
5. Singer, S.F. Meteorological measurements from a minimum satellite vehicle. *Trans. Am. Geophys. Union* 38 (4): 469–482 (1957).
6. Rao, P.K. (ed.). *Weather Satellites: Systems, Data, and Environmental Applications*. American Meteorological Society, Boston, 1990.
7. Rao, P.K. Evolution of the Weather Satellite Program in the U.S. Department of Commerce: A Brief Outline. NOAA Technical Report NESDIS 101, Washington, DC, 2001.
8. Rao, P.K., and J.S. Winston. An investigation of some synoptic capabilities of atmospheric ‘window’ measurements from satellite TIROS-2. *J. Appl. Meteorol.* 2 (1): 12–23 (1963).
9. Winston, J.S., and P.K. Rao. Temporal and spatial variations in the planetary-scale outgoing long-wave radiation as derived from TIROS-2 Measurements. *Mon. Weather Rev.* 641–657, (Oct–Dec 1963).
10. Singer, S.F., et al. Meteorological satellites: Technology and applications. *Astronaut. Aerospace Eng.* 1: 61–66 and 89–92 (April 1963).
11. Yates, H., et al. Terrestrial observations from NOAA operational satellites. *Science* 231: 463–470 (1986).

12. Chapman, R. *TIROS-Nimbus: Administrative, Political, and Technological Problems of Developing U.S. Weather Satellite*. The Inter-University Case Program, Inc. Syracuse, NY, 1969.
13. Kennedy, J.F. (President of the United States). Urgent National Needs. Message to the Congress, May 25, 1961 (H-Doc. No. 174, 87th Congress, 1st Session).
14. Smith, W.L., et al. The meteorological satellite: Overview of 25 years of operation. *Science* 231: 455–462 (1986).
15. Widger, W.K. Jr. *Meteorological Satellites*. Holt, Reinhart & Winston, New York, 1966.
16. Vaughn, W.A. Satellites: Past, Present, and Future. NASA Conference Publication 2227, May 1982.
17. Chen, H.S. *Space Remote Sensing Systems: An Introduction*. Academic Press, New York, 1985.
18. Curran, P.J. *Principles of Remote Sensing*, Longman Group, London, 1985.
19. Goody, R.M., and Y.L. Yung. *Atmospheric Radiation: Theoretical Basis*. Oxford University Press, Oxford, 1989.
20. Carleton, A.M. *Satellite Remote Sensing in Climatology*. CRC Press, Boca Raton, FL, 1991.
21. Reeves, R.G. (ed.). *Manual of Remote Sensing*, Vols. 1 & 2. American Society of Photogrammetry, Falls Church, VA, 1975.
22. Asner, G.P., and R.O. Green. Imaging spectroscopy measures desertification in United States and Argentina. *Eos* 82: 601 (4 Dec. 2001).
23. Singer, S.F., and F. Williams. First observations of sea state based on microwave emission. *J. Geophys. Res.* 73: 3324–3327 (1968).
24. Grody, N.C. Classification of snow cover and precipitation using the special sensor microwave/imager (SSM/I). *J. Geophys. Res.* 38: 7423–7435 (1991).
25. Spencer, R.W., et al. Heavy thunderstorms observed over land by the Nimbus-7 scanning multichannel microwave radiometer. *Appl. Meteor.* 22: 1041–1046 (1983).
26. Spencer, R.W., and J.R. Christy. Precision and radiosonde validation of satellite gridpoint temperature anomalies. Part 1: MSU Channel 2. *J. Climate* 5: 847–857 (1992).
27. Svejksky, J., and B. Jones. Satellite imagery detects coastal stormwater and sewage runoff. *Eos* 82: 621 (11 Dec. 2001).
28. Long, D.G., et al. Global ice and land climate studies using scatterometer image data. *Eos* 82: 503 (23 Oct. 2001).
29. Drinkwater, M.R., and C.C. Lin. Introduction to the special section on emerging scatterometer applications. *IEEE Trans. Geosci. Remote Sens.* 38: 1763–1764, 2000.
30. Singer, S.F. Measurement of atmospheric surface pressure with a satellite-borne laser. *Appl. Opt.* 7: 1125–1127 (1968).
31. Fritz, A., and H. Wexler. Cloud pictures from TIROS-1. *Mon. Weather Rev.* 88: 79–87 (1960).
32. Wexler, H., and S. Fritz, TIROS reveals cloud formations. *Science* 1–131 (3415): 1708–1710 (1960).
33. Oliver, V.J. TIROS pictures a pacific frontal storm. *Weatherwise* 13 (5): 186 (1960).
34. Rossow, W., and P.A. Schiffer. Advances in understanding clouds from ISCPP. *Bull. Am. Meteorol. Soc.* 80: 2261–2287 (1999).
35. Stephens, G.L., G.G. Campbell, and T.H. Vonder Harr. Earth radiation budget measurements from satellites and their interpretation for climate modeling and studies. *J. Geophys. Res.* 86: 9739–9760 (1981).
36. House, F.B., A. Gruber, G.E. Hunt, and A.T. Mecherikunnel. History of satellite missions and measurements of the Earth radiation budget. *Rev. Geophys.* 24: 357–377 (1986).
37. IPCC. *Climate Change*. Cambridge University Press, Cambridge, 2001.

38. National Research Council. *Reconciling Observations of Global Temperature Change*. National Academy Press, Washington, DC, 2000.
39. National Research Council. *Toward a New Weather Service, Continuity of NOAA Satellites*, National Academy Press, Washington, DC, 1997, p. 51.
40. Iannotta, B. New eyes on Earth's weather. *Aerosp. Am.* 39: 33–39 (2001).

S. FRED SINGER

The Science & Environmental Policy Project (SEPP)
Arlington, Virginia

P. KRISHNA RAO

National Environmental Satellite
Data, and Information Service
National Oceanic and Atmospheric Administration
Silver Spring, Maryland