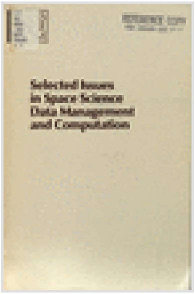


Selected Issues
in Space Science
Data Management
and Computation

Selected Issues in Space Science Data Management and Computation (1988)

Pages
76

Size
5 x 9

ISBN
0309319439

Committee on Data Management and Computation; Space Science Board; Commission on Physical Sciences, Mathematics, and Resources; National Research Council

Selected Issues in Space Science Data Management and Computation

Committee on Data Management and Computation
Space Science Board
Commission on Physical Sciences, Mathematics,
and Resources
National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1988

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organisation of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognises the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organised by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

Support for this project was provided by Contract NASW 3482 between the National Academy of Sciences and the National Aeronautics and Space Administration.

Available from
Space Science Board
2101 Constitution Ave., N.W.
Washington, D.C. 20418

Printed in the United States of America

COMMITTEE ON DATA MANAGEMENT AND COMPUTATION

Christopher T. Russell, University of California, *Chairman*
Mark Abbott, Scripps Institution of Oceanography
Ted Albert, U.S. Geological Survey
Lawrence K. Bolef, University of Colorado
Stephen Brecht, Berkeley Research Associates
John E. Estes, University of California
Giuseppina Fabbiano, Center for Astrophysics
Owen K. Garriott, Teledyne Brown Engineering
John C. Gille, National Center for Atmospheric Research
Elaine Hansen, University of Colorado
Harold Masursky, U.S. Geological Survey
Lucy McFadden, University of California, San Diego
Nicholas Roussopoulos, University of Maryland
Peter Shames, Space Telescope Science Institute

Daniel N. Baker, Goddard Space Flight Center, *Liaison Representative*

Richard C. Hart, *Executive Secretary*
Carmela J. Chamberlain, *Administrative Secretary*

SPACE SCIENCE BOARD

Thomas M. Donahue, University of Michigan, *Chairman*
D. James Baker, Joint Oceanographic Institutions, Inc.
Roger D. Blandford, California Institute of Technology
Larry Esposito, University of Colorado
Lennard Fisk, University of New Hampshire
Jonathan Grindlay, Center for Astrophysics
Donald Hall, University of Hawaii
William Kaula, NOAA
Harold Klein, The University of Santa Clara
John Leibacher, National Solar Observatory
Robert Pepin, University of Minnesota
Roger Phillips, Southern Methodist University
David Raup, University of Chicago
Christopher Russell, University of California, Los Angeles
Blair D. Savage, University of Wisconsin
John Simpson, Enrico Fermi Institute, University of Chicago
George L. Siscoe, University of California, Los Angeles
L. Dennis Smith, Purdue University
Darrell Strobel, Johns Hopkins University

Dean P. Kastel, *Staff Director*
Richard C. Hart, *Senior Program Officer*
Joyce M. Purcell, *Senior Program Officer*
Paul F. Uhlir, *Senior Program Officer*

COMMISSION ON PHYSICAL SCIENCES, MATHEMATICS, AND RESOURCES

Norman Hackerman, Robert A. Welch Foundation, *Chairman*
George F. Carrier, Harvard University
Dean E. Eastman, IBM, T.J. Watson Research Center
Marye Anne Fox, University of Texas
Gerhart Friedlander, Brookhaven National Laboratory
Lawrence W. Funkhouser, Chevron Corporation (retired)
Phillip A. Griffiths, Duke University
J. Ross Macdonald, University of North Carolina, Chapel Hill
Charles J. Mankin, Oklahoma Geological Survey
Perry L. McCarty, Stanford University
Jack E. Oliver, Cornell University
Jeremiah P. Ostriker, Princeton University Observatory
William D. Phillips, Mallinckrodt, Inc.
Denis J. Prager, MacArthur Foundation
David M. Raup, University of Chicago
Richard J. Reed, University of Washington
Robert E. Sievers, University of Colorado
Larry L. Smarr, National Center for Supercomputing Applications
Edward C. Stone, Jr., California Institute of Technology
Karl K. Turekian, Yale University
George W. Wetherill, Carnegie Institution of Washington
Irving Wladawsky-Berger, IBM, Data Systems Division

Raphael G. Kasper, *Executive Director*
Lawrence E. McCray, *Associate Executive Director*

Preface

The Space Science Board's Committee on Data Management and Computation (CODMAC) has published two reports aimed at improving the scientific return from data acquired by space missions. The first of these, *Data Management and Computation, Volume 1: Issues and Recommendations* (National Academy Press, 1982), hereafter referred to as CODMAC I, summarized the major problems that have been impediments to extraction of science information from space-acquired data, recommended a number of general steps for improvement, and developed a set of principles for successful management of scientific data. The second report, *Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences* (National Academy Press, 1986), hereafter referred to as CODMAC II, explored management approaches and technology developments for computation and data management systems designed to meet future needs in the space sciences. This report continues these studies by examining several specific areas where improvements in NASA's data management and computational resources could be of substantial benefit to science.

On behalf of the committee, I would like to thank the following for assisting us in preparing these white papers and/or for participating at our study in Snowmass, Colorado, in August 1986 when we first drafted the documents: Dave Nichols, John Solomon, and Tom Duxbury (Jet Propulsion Laboratory); Carroll Hood (Science Applications International Corporation); Mike Wiskerchen (Stanford University); George Ludwig (University of Colorado); Joe King and Phil Cressy (Goddard Space Flight Center); Ray Walker (University of California, Los Angeles); and Ethan Schreier (Space Telescope Science Institute).

Christopher T. Russell
Chairman

1
Introduction

Scientific advances arising from data acquired by NASA spacecraft require more than just the data themselves. Also required are the means to save and later access the data, computers to reduce the data and create physical models, communications to move the data from one place to another, and structures to manage the data and associated resources.

Committee on Data Management and Computation (CODMAC) has studied many of these areas of data management and computation previously. The result was two reports that made recommendations for improving the scientific return from space-acquired data. After the second report appeared in early 1986, the committee decided to address a number of specific issues in more depth. The findings are summarized in chapter 2. The succeeding chapters provide supporting details.

Some of the specific issues CODMAC chose to examine arose because of the implementation of new technologies, such as networking to utilize computer power and databases in a more cost-effective manner. NASA's development of the Program Support Communications Network, the success of the Space Physics Analysis Network (SPAN), the growth of systems such as ARPANET, and the establishment of an NSF supercomputer network are examples of the recognition of networks as a fundamental tool. The astronomical growth of electronic mail services indicates that network use is rapidly becoming an essential feature of modern society. In chapter 3, "Computer Networking for the NASA Scientific Community" (principal authors: D.N. Baker, T. Duxbury, E. Schreier, P. Shames) the committee addresses a number of issues in this area and offers NASA several recommendations for ways to proceed with networks for scientific use.

Supercomputing has become, in the past few years, an effective and efficient tool for doing science. Space missions, for example, can be optimized by first modeling space systems (such as a planetary magnetosphere or a comet) before building an expensive probe to go there. NSF clearly recognized this trend in their program to establish supercomputer centers for use by the scientific community. Within NASA, there is a growing trend at the centers to acquire supercomputers. Unfortunately, these machines are not usually accessible for scientific purposes by other than center scientists, and even then access may be difficult. In chapter 4, "Supercomputing for NASA Funded Sciences: Resources and Access" (principal authors: S. Brecht, R. Walker, C. Hood), some serious problems are pointed out in NASA's approach to utilizing supercomputers for science.

In chapter 5, "The Management of High Data Rates and High Data Volumes" (principal authors: M. Abbott, D. Nichols, J. Solomon), ways are suggested for NASA to deal with the expected volume of data from future space programs. We are now able to generate data at rates well beyond our ability to digest. Such data streams must be harnessed for efficient scientific analysis if the ready availability and cost-effective management of these data are to be realized.

In chapter 6, "NASA Data Management Issues" (principal author: T. Albert), the focus returns to a constant CODMAC concern—the necessity for proper data archiving, and effective access and distribution policies. There are vast archives of data that are not well used. We must find ways to preserve and nourish this vital resource.

This report does not address the growing desire on the part of the scientific community for real-time or near-real-time access to the scientific telemetry stream. This desire has been expressed for small missions, such as Explorer-class missions, for Observatory-class missions, and for Space Station-based investigations. This issue is a complex one involving issues of the security of the communication links and the cost/benefit ratio of each application. This issue will be left for a future study.

2
Executive Summary

COMPUTER NETWORKING FOR THE NASA SCIENTIFIC COMMUNITY

A successful computer network must provide mail, remote logon, and file transfer capabilities. For maximum utility it must be connected to a large number of sites. It should be reliable and have high bandwidth and low cost. Several different network protocols exist at the present time that support successful networks. Moreover, international network standards are evolving that may require changes in the future networks. Thus, at present there is not a single networking protocol that can be recommended for all users, and care must be exercised to allow for future changes. In this environment CODMAC makes the following recommendations.

1. *Continue support of SPAN in its present form.*
2. *Implement a TCP/IP-based NASA science network to support high data rate science and broad community access and connect this network to similar networks of other agencies.*
3. *Continue to support at least two major network protocols such as DECNET and TCP/IP.*
4. *Connect these NASA-supported networks to NASA's major computational facilities and data archives.*

5. Develop and support data description languages, formatting standards, data compression techniques, and naming and documenting conventions that will encourage broad access to a wide range of data sets.

6. Ensure that the applications software and interfaces developed will survive the transition from current protocols to ISO/OSI standards.

7. Establish a NASA-wide, coherent, electronic mail system with guaranteed delivery.

8. Provide adequate support for the NSI, the PSCN backbone, and tail-circuits.

9. Encourage strong and continuing user involvement in the development, evolution, and operations of the network.

SUPERCOMPUTING FOR NASA FUNDED SCIENCES: RESOURCES AND ACCESS

Computers are now pervasive in all aspects of science from the collection of data to its analysis. The ever increasing bandwidth of our sensors and the ever increasing sophistication of our models and analyses place ever increasing demands on our computational resources. We expect that there will always be a demand within the NASA science community for the largest and most powerful computers (i.e., supercomputers) no matter how sophisticated and fast smaller computers become, because larger computers will become commensurately faster and their memories larger. NASA has entered the supercomputer era and now possesses nine supercomputers at its various centers. While this may appear to provide an abundance of supercomputer resources, there is no way at present to gauge the availability of these resources or to allocate them if their availability is ascertained. CODMAC therefore makes the following recommendations.

1. NASA should develop an agency-wide plan for the management of its supercomputer resources.

2. OSSA should prepare a strategic plan for supercomputing for NASA space science users.

3. The OSSA centers should recognize and anticipate the needs of their user communities and be an advocate for them when procuring resources for scientific computing.

4. Users of supercomputing should be involved at every level in the development of the supercomputing plans and should participate in the oversight of the operations of NASA's supercomputing resources.

THE MANAGEMENT OF HIGH DATA RATES AND HIGH DATA VOLUMES

The high data rates and volumes planned for the sensors and communication systems of the Space Station era will present challenges for efficient data acquisition, processing, and management. A strategy must be developed to provide the greatest scientific return within the given resources. To do this, scientific users must have access to appropriate tools and techniques. There must also be an aggressive technology development program for high rate/high volume data and information systems. CODMAC's recommendations for addressing these issues are as follows:

1. The goal of any data management strategy must be to maximize the scientific return from the acquired data within the constraints of the data system. Users must be given an active role in the planning of data collection both to ensure high scientific return from the data and to increase awareness in the users of the implications of various observing scenarios.

2. Future space science missions should develop a data acquisition strategy that allows the user to participate in the data management process. This participation should include consideration of total life-cycle costs, interactive payload control to allow data editing based on quality, and mechansims for conflict resolution.

3. OSSA must aggressively pursue the development of tools and techniques that will enable a robust data rate management strategy to be adopted. Performance models should be developed. Testbeds of methods for on-board information extraction and autonomous instrument control should be implemented and data compression techniques should continue to be pursued.

NASA DATA MANAGEMENT ISSUES

While the collection of data from space attracts much interest, the more mundane issues on handling, managing, and disseminating these data are given less than adequate attention. The accumulation of data continues as do the data management problems with the loss of valuable data and with the difficulty of providing access to the data. There is too much data to be efficiently collected and stored in a single location. Moreover, data is best stored where specialists who use the data and know the most about it reside. If potential users are to gain access to the data, good directories and catalogs must exist. If the data are to be adequately preserved they must be stored on suitable media. In an era of evolving technology, assessing the quality of the available media becomes critical to the preservation of the archive. In view of the present situation and recognizing that some strides in these directions have been made, CODMAC makes the following recommendations.

1. NASA should adopt and implement an explicit data management plan for all space flight investigations.

2. NASA should provide sufficient resources for data archiving, guidelines for its implementation and enforce the requirements that projects and principal investigators properly archive and document their data.

3. NASA should develop procedures for the protection of the data archive from deterioration of media, hardware failures, and tampering by individuals.

4. There should be an active, distributed archive managed in scientific data management units by each discipline in coordination with NSSDC.

5. NASA should develop easily accessible, standard catalogs and directories.

6. NASA should continue to assess storage media and develop guidelines for its use in archiving.

7. NASA should establish an advisory committee on data retention and preservation and associated concerns, and should establish ties with other agencies and the user community regarding the dissemination of and access to archived data.

8. NASA should support and promote the use of its data archives.

3
Computer Networking for the NASA Scientific Community

SUMMARY

Committee on Data Management and Computation has examined the outstanding issues related to NASA computer networking and reviewed the ongoing networking efforts in the light of the principles previously set down by CODMAC for data systems. Based on this examination and review, the committee recommends that NASA should:

1. continue support of SPAN in its present form;

2. implement a TCP/IP-based NASA science network to support high data rate science and broad community access and connect this network to similar networks of other agencies;

3. continue to support at least two major network protocols such as DECNET and TCP/IP;

4. connect these NASA-supported networks to NASA's major computational facilities and data archives;

5. develop and support data description languages, formatting standards, data compression techniques, and naming and documenting conventions that will encourage broad access to a wide range of data sets;

6. ensure that the applications software and interfaces developed will survive the transition from current protocols to ISO/OSI standards;

7. establish a NASA-wide, coherent, electronic mail system with guaranteed delivery;

8. provide adequate support for the NSI, the PSCN backbone, and tail circuits; and

9. encourage strong and continued user involvement in the development, evolution, and operations of the network.

BACKGROUND

The numerous benefits of computer networks for the scientific community are well documented (cf., CODMAC II). Wide-area networks are crucial to developing distributed information systems and providing connection links between Space Science Data Management Units (SSDMUs). In serving this function, networks enhance communications and productivity and make data and computational resources accessible to the broadest elements of the scientific community in a fast and cost effective manner.

The most immediate and basic requirements for a successful computing network are that it provide mail, remote log-on, and file transfer capabilities for a distributed user community with a variety of different computers. In many respects, it is not important to the majority of the network users which particular communication protocols are used, nor is the detail of the underlying system configuration of particular concern. For most scientific users the structure and inner functioning of the network should be as transparent as possible. The most important attributes to the intensive user are the number of sites that are connected to the network (its connectivity), the operations that can be carried out over the network (its functionality), and its speed, cost, and reliability (its efficiency). This chapter explores several issues relating to space science networks and their protocols, configurations, and management.

ISSUES: THE NEED FOR MULTIPLE NETWORKS

A satisfactory way to gauge the requirements for a computing network is to identify its most taxing activities and functions. The network must then be designed in such a way that its bandwidth and reliability can readily meet the requirements of these most difficult functions. Obviously, a network so configured will be able to deal effectively with the less demanding tasks that are part of its overall operational profile. Part of this design activity is to understand both short-term peaks in usage profiles and longer-term trends driven by changing requirements in data and connectivity.

In the next decade, the committee envisions that information processing associated with Space Station activities will be among the most stressing of the requirements faced within the space science community. Those undertaking investigations on the Space Station will wish to network electronically during the experiment test and phase, to command and operate their instruments remotely and analyze the resulting data remotely, and to communicate with other users. Any computer network that meets the user access requirements underlying the goals in the space science era (by exhibiting adequate reliability, connectivity, functionality, and communication bandwidth) will also meet the broader needs of the space science community vis-a-vis most computer networking objectives. There are some specific requirements, i.e., for bandwidth or privacy, that may need special accommodation for these requirements to be met—remote image display and remote display of supercomputer-generated models are two current examples.

At the present time there are many separate scientific networks in the United States, supported both by national agencies and as "grass roots" efforts. These range from relatively small networks sending limited volumes of data (bit streams) to very large, high-volume networks with hundreds of users. Networks to support NASA science users do not exist in a vacuum, since several other national agencies and consortia already have, or are putting, networks in place. The oldest of these is ARPANET, which has been in existence since 1969 and is the conceptual parent of all of today's major networks. ARPANET and the extensions being built by NSF and various university and regional consorts use the Transmission Control Protocol/Internet Protocol (TCP/IP) protocols, which are widely supported on a variety of machines from the micro to the supercomputer class.

Significant networks that use other proprietary protocols such as DECNET (DEC) and SNA (IBM) also exist and support extended communities of users. Protocols that offer only mail service are also in wide use, with BITNET, USENET, and CSNET being the most common in academia and Telenet finding wide use in NASA. Public data nets using the X.25 protocols (which support Telemail) are in wide use in the United States and are often the only network in many other countries, providing ubiquitous, if slow, access to host systems. For the purposes of this report, we will restrict our attention to two basic "full function" networks that have found wide acceptance in the NASA user community. The first of these networks is SPAN, which is a DECNET-protocol system. SPAN serves several hundreds of users and links together several tens of space physics institutions; SPAN has evolved to use the Project Support Communications Network (PSCN) backbone plus a router/tail circuit configuration. The second network is a proposed NASA Science Network (NSN), which will be very similar to the ARPA and NSF networks in configuration and will use TCP/IP communication protocols. The NSN concept is evolving out of various pilot program requirements, primarily the Planetary Land Data System (PLDS) and the Astronomy and Astrophysics Network Project—currently in the process of connecting a dozen major astronomy centers together via ARPANET until NSN is implemented.

The various NASA networks must be interconnected with each other and with other agency networks. To implement this a NASA Science Internet (NSI) has been proposed. This internet not only should tie together the various TCP/IP networks (such as NSN, ARPANET, and NSFNET), but also should provide connectivity with networks of differing protocols (such as SPAN) as appropriate and where feasible.

In the appendix, we give more complete descriptions of the history, configuration, and management plans for the SPAN and NSN systems. Suffice it to say here that with early and continuing user involvement, these networks meet, or will meet, the requirements of a large fraction of the NASA science and technology communities today. These networks will evolve (as will be discussed below) to meet future needs, including international standard protocols when these become established. In all of our discussion of specific networking issues, we suggest that the reader keep the SPAN and NSN models in mind as concrete, prototypical examples of what a space science computing network can and will be.

Because of the extensive networks being put in place by other agencies, it is strongly suggested that NASA reach cooperative agreements with NSF and DARPA to share technology and network access. A national initiative under the Federal Coordinating Committee on Science and Engineering Technology (FCCSET) sponsorship to consider an Interagency Research Internet is now in progress, and NASA should give serious consideration to this effort. The particular needs of NASA for high-bandwidth, remote data access can best be met by a NASA network, but access to the broad community of scientists, researchers, other government agencies, and parts of the commercial sector will be well met by interagency network arrangements.

Recommendation

CODMAC recommends the development of a NASA Science Network to meet the present need for network services employing the TCP/IP protocol. CODMAC also recommends that NASA continue to support SPAN in its current form. These two networks should be interconnected and also connected to similar networks of other agencies. For the foreseeable future, NASA should continue to support at least two major network protocols such as DECNET and TCP/IP.

User Involvement in Network Operations and Planning

Present and future NASA computing networks exist to serve the scientific and technical communities. The networks are not an end unto themselves. The only way to ensure continued relevance and responsiveness of any network is to have a strong involvement of the scientific users in the management, operation, and configuration of the network.

It is, of course, clear that a staff of professionals should carry out the hands-on, day-to-day operation of a network. This staff can bring to these operations the best and the most informed decisions about hardware and other technical matters. But it is the scientific user of a network who understands what is being accomplished with the network and/or what needs to be improved to achieve the desired goals. To this end, the users must have an organized role in the management of network activities.

The SPAN system can be viewed as something of a model for user involvement in a scientific network. In its day-to-day operation, SPAN is run by computer professionals within the National Space Science Data Center (NSSDC) at Goddard Space Flight Center (GSFC) and at Marshall Space Flight Center (MSFC). However, approximately twice each year a representative cross section of involved users of SPAN, the Data Systems Users Working

Group (DSUWG), meets for 2 or 3 days for information exchange and to discuss operation, development plans, and standards issues as they affect the network. A given DSUWG meeting may involve 25 to 30 percent of the active SPAN sites and is an open forum for discussion of problems and promise. A small (5-member) steering committee of the DSUWG executives is charged with much more frequent interaction with the SPAN operational staff, and the steering committee provides continuous user advise and consent to SPAN operations between major DSUWG meetings.

User groups obviously can play a strong role in defining initial network configurations and functionalities. Moreover, any reconfigurations or additions to networks must take user wishes into account. User groups can and should help reduce overall NASA costs for networking by optimizing system performance and system configuration. Elimination of redundant tail circuits, low-traffic lines, and unnecessary nodes and aggregation of several low-speed circuits into a single higher speed one can greatly reduce Agency costs, but these decisions can only be made intelligently by the users themselves. The recent Information Systems Office (ISO) establishment of a PSCN User's Committee to provide a focus for science user requirements presents a clear example of this, where users from disparate disciplines found they had very similar requirements, and had overlapping requests submitted to the PSCN. The resulting coordination within the PSCN user community and between PSCN users, NSF, and ARPANET users represents a potentially great decrease in cost and increase in functionality and connectivity.

As networks evolve and grow, the most successful of them will undoubtedly expand to involve many different disciplines or subdisciplines. This breadth of coverage provides a valuable economy of scale and will ensure the widest possible interconnection of SSDMUs, institutions and users. We encourage, however, the continued existence within such networks of user groups representative of different disciplines. Individual user groups with common requirements and goals from "communities of interest" within the larger network may be treated as logical entities with various network services provided they are tailored for their needs. These groups should have active representation on any steering committee so that the network remains responsive to their needs.

Recommendation

NASA should encourage strong and continuing involvement of network users in the development, evolution, and operations of the network to ensure relevance, responsiveness, and efficiency of network resource utilization.

Access to Other Major Networks

It is clear that computer networks are being developed to serve virtually all scientific disciplines in the United States, and furthermore, numerous networks are being developed in Europe, Canada, Japan, and elsewhere. These networks serve a variety of purposes and users. They are often based on different communication protocols and they entail several different architectural concepts. Because of the growing commonality of scientific interests and/or the need to share major resources and databases, there is an increasing need to provide connectivity between networks. Connections between major networks that share a common set of protocols can be effected through use of gateway machines. Connections between heterogenous networks can be provided by "Janus Hosts" or application level gateways that can bridge across disparate protocols. They represent an effective technique but can also be costly, and a high overhead can be paid in terms of system efficiencies. This can limit substantially the transfer capabilities of such a hybrid network—but does provide wider connectivity and a higher level of interoperability than can otherwise be achieved.

For networks sharing common protocols and architectures, it often may be advisable to effectively merge the networks into one larger entity or internet. Depending on the protocols in use, this may be a simple operation but can be a formidable problem when it comes to node naming conventions and other management activities in the larger system. Resolution of such conflicts can absorb substantial time and attention of system managers. Central management of system and host names and ID numbers is essential. The TCP/IP-based networks recognized this early in their development, and they have run for some time with centralized name and number management.

A matter of concern as network links are established is the controlling of access of external users to the resources and databases that may reside on a given network. NASA, for example, may be concerned about a European Space Agency (ESA) network user accessing mission data that reside on SPAN or NSN. Similarly ESA may be sensitive to uncontrolled access by U.S. researchers of ESA data sets or resources. NASA managers, site managers, and network operations personnel must pay particularly close attention to these issues of network-to-network connection and host access, and NASA must provide some institutional policy and guidelines of permissible access.

Notwithstanding the difficulties that may result when networks touch one another, the linkage with major networks such as NSFNET, ARPANET, etc., is crucially important if NASA networks are to provide full and proper functionality. Therefore, NASA must plan explicitly for establishing contact between major U.S. and scientific networks. Moreover, substantial effort and attention should be devoted to establishing effective links with European networks (e.g., ESA) as are appropriate to the NASA disciplines and with emerging Japanese, Canadian, and other international networks.

Existing and planned network connections must support at least two protocols—TCP/IP and DECNET—and the future goal of supporting the ISO/OSI suite has also been established within NASA. It is clear that this last objective will be significant for future European and other foreign connections. However, it should also be noted that the realizations of networks based on these truly standard protocols is several years in the future.

Recommendation

Because it is important to ensure availability of various science resources to the widest community, we strongly encourage all major NASA resource sites to connect to at least the two major network families and to provide services equally to both. As new network access services are provided, we strongly encourage use of protocol layers that will encourage interoperability and will permit easy transition to the ISO/OSI networks as they become available.

Network Connections for Resource Usage

In addition to the generic "full service" network capabilities, consisting of mail, remote log-on, and file transfer, there are specialized resources that NASA's space science community must be able to access via networks. These resources fall in three general categories: supercomputing; data archives, catalogs, software libraries, and other databases; and remote observing and related instrument activities. Each of these may impose performance constraints on the generic network capabilities or may require specialized capabihties. These needs have been recognized in the planning for NSI and testbed activities are expected to explore the options for fulfilling these requirements.

Supercomputing

Use of supercomputers via a network generally imposes significant demands on the network beyond remote log-on. When supercomputers are used for data analysis, there is a significant amount of information that must be processed, e.g., images or synthetic aperture radar (SAR) data. The amounts of data involved (megabytes to gigabytes) are such as to overload any presently achievable network. The current solution will likely involve setting up procedures whereby the data will be sent to the supercomputer nonelectronically (e.g., via optical disks or tapes), and the codes and interaction with the running program will be transmitted over the network.

The transmission of the output of supercomputing is a less tractable problem. Such output is, again, often in the form of images or graphics. It should be remembered that, especially in theoretical and modeling studies, supercomputers can be prodigious generators of data. The most promising approaches currently being considered involve using graphics/image processing workstations, interfaced to the supercomputer via the network. The output would be transmitted as compressed data or as high level graphics instructions, and the local work station would reconstruct the image. The supercomputer and the network are then relieved of the job of image display and manipulation. Although these ideas are accepted in principle, and work is proceeding in the research and commercial sectors, there are as yet no widely accepted protocols for remote workstation-supercomputer links and the attendant data compression and graphics/image processing standards.

Data Archives and Catalogs

The importance of data archives and remote access to them was stressed at length in both of the previous CODMAC reports. The aspects of this subject that are most relevant to the question of networks are access to catalogs and data browsing. First, as stressed in CODMAC II, there must be a well-publicized, easy to access central directory describing the various space science archives and explaining how to access them—what networks they are on and "how to get in the front door." This task has been undertaken by the NSSDC and must be made easily accessible on all NASA-supported networks. Each of the active archives must maintain its own catalog; the existence of networks and remote access makes it less important to have all the catalogs in the same place and yet provides the means to coordinate access.

Since the types of data and the structures of the archives may be radically different, it should be expected that catalogs will be structured differently. Thus, each catalog must be self-documenting, but certain basic query functions must be supported, and we would encourage development and adoption of a common data definition language capable of describing the wide variety of present and future data sets. The concept of distributed database management systems is currently being studied actively. We can look forward to networked systems where multiple database structures and multiple query languages can coexist, using standard interfaces. We encourage NASA to stay abreast of this technology in the development of new archives—e.g., NSSDC, Planetary Data System (PDS), Hubble Space Telescope (HST).

Once data are located in an archive, it should be possible to selectively view them via the network. This so-called browse capability allows a user to verify the utility of the data before requesting potentially massive amounts. It should be remembered that current networks are not appropriate for transmitting large amounts of data; the concept of browsing involves transmitting some subset of the data, for example, a reduced dynamic range or reduced spatial resolution image. Some techniques for browsing are likely to be very similar to those discussed for supercomputing: access by workstations via networks using standard data compression techniques and high-level graphics instructions. However, there are also likely to be sampling techniques appropriate to a given discipline. It will be essential for archival catalogs to document these browse mechanisms; it should also be possible to download any relevant algorithms to the user facility for decoding the browse data.

This last point is related to the larger issue of software coordination. As was pointed out in CODMAC II, data centers will be called on more and more to support software libraries appropriate to a given discipline. This software, necessary to analyze the data archives, will also be available via the network. As extensive portable software systems become more widespread, data centers should be prepared to maintain these systems and to distribute updates via the network as appropriate.

Data archives and catalog access are important in support of several scientific goals. Beyond direct access to single data sets, there is also the desire to support various coordination activities. Coordinated workshops where users come together to work jointly on a single problem is one example. Another example is use of network access to perform correlative studies across several data sets that are distributed in different archives.

Remote Observing/Operations

Many aspects of experiment operations can be done via networks. These range from mission planning and scheduling, through real-time control and monitoring, to data acquisition. There are many advantages, including decentralizing operations, decreasing transcription errors, providing more rapid turn around, and, of course, cutting travel costs. Several examples of such concepts can be mentioned. The SPAN support of the ICE comet rendezvous demonstrated rapid international coordination of real-time data analysis. The Space Telescope Science Institute (STScI) currently supports submission of observing proposals via TELENET. This greatly decreases the possibility of errors being introduced in the process of data entry: not only is there direct machine-to-machine transfer, but on-line validity checking gives the user immediate feedback as to syntactical or consistency errors. This system will be expanded to include the catalog of accepted proposals, and will be accessible via the astronomy network in the near future. The Mars Observer project is planning distributed mission planning and scheduling, with each investigator having access to scheduling workstations.

Remote observing is being tried in ground-based observatories. Observing commands are relayed to on-site operators. Such experiments are limited by lack of reliable networks and of standard techniques for transmitting final images and the equivalent of quick-look data. Again, the same techniques and protocols for transmitting compressed data and image/graphics instructions, as discussed for supercomputing and archive browsing, will be relevant to remote observing. Coordinated observing at ground-based and space-based facilities is another interesting possibility that will be facilitated by broad connectivity to a wide community.

Recommendation

NASA should support the provision of supercomputing and data archives on its networks and should encourage remote experiment observations and operations via present and future network systems. This includes the development and support of protocols and access services, data description languages, formatting standards, data compression techniques, and documenting conventions that will encourage and allow broad access to a wide range of data sets and facilities.

Ease of Use

The prime consideration for ease of use for a network is transparency. A user should not need to know the technical details and physical topology of the network he or she is using. Nor should the user need to understand the details and topologies of other networks to which colleagues are connected, or the gateways between the networks. Finally, users should not need to know the intricacies of particular hardware/software systems that may be vendor specific. To quote from the statement of work for the NASA Science Network, a network should "provide a common set of user network access functions that . . . are independent of underlying, vendor particular hardware or software systems to provide interoperability in the existing heterogeneous computer system environment within the science community."

An example of user-friendliness in networking is user directories. One should only need to know a user/node name and perhaps a supporting identifier indicating discipline domain. The network services ought to be able to access the directories and supply the necessary routing information to connect one user to another. One can imagine each discipline logical network, or community of interest, having not only its own user directory, but also a directory of services available, and instructions on how to access these services.

Another example of user-friendliness involves mail. At the current time, there are many parallel networks supporting different user communities. Thus, a person might have to access numerous electronic mailboxes each day and learn how to route across multiple gateways. Although it is possible to supply site-specific procedures to access these multiple mailboxes and gateways, inform the user the mail exists, etc., it is clear that standard network services ought to include automatic relay services, whereby interfaces between, e.g., TELEMAIL, BITNET, DECNET, ARPANET, and UUCP are automatically established. And to be useful, mail delivery ought to be guaranteed, not just broadcast, with the hope that all the links are in place.

A final, very specific example involves NASA's electronic mail itself. All members of the NASA family (e.g., agency personnel, university scientists, industry contractors, and advisory committees) require straightforward, simple, and rapid means of communication. A coherent electronic mail service is essential. We are concerned that the decentralization of NASA's electronic mail system has taken place without prior discussion with the user community and without consideration of ease of use. Thus, it has become much harder to maintain electronic communication among groups that span different NASA centers.

Recommendation

NASA should establish an agency-wide and coherent electronic mail system with guaranteed delivery and gateways to all the widely used public and commercial data networks.

Network Evolution and Standardization

There are three major issues related to network evolution and standardization: (1) consolidation of current and proposed networks adopting the same protocol so that they can all communicate with one another; (2) the development of interconnecting gateways and relays between networks to increase connectivity and communication; and (3) the evolution of network protocol to the ISO/OSI standards to provide the universal ability to communicate between networks.

Currently, the Office of Space Science Applications (OSSA) of NASA is proposing separate networks to support the astrophysics community (Astronet), the land community (Earth Science Network), the planetary community (planetary Data Network), etc., all using the TCP/IP protocol. The thrust of the NASA Science Network within the PSCN to consolidate these networks and extend support to additional science disciplines is encouraged. This will guarantee the ability to communicate between disciplines while offering networking services to a much broader community. It is recognized that mechanisms will be needed to ensure the integrity of each discipline's requirements, and strong user involvement.

There are existing and planned networks outside of NASA that will support a large community that also contains NASA scientists. The NSFNET, ARPANET, and a planned NOAA Network are prime examples. These networks will connect user institutes, data archives, observing sites, research laboratories, and computer facilities that are important to the NASA science community. Therefore, gateways between the NASA networks and these other networks are crucial for NASA to fulfill its science role, and agreements need to be developed for the sharing of resources between these agencies. The NASA Science Internet is working to establish linkages between these TCP/IP based communities and those running the DECNET protocols. The aim is to achieve the greatest amount of interconnections and interoperability.

It is felt that network hardware and software that fully implements the ISO/OSI protocol standards will be available from commercial vendors within a few years. All NASA networks should maintain a development path that will allow them to evolve to these ISO protocol standards when available. Choosing this evolutionary track will keep NASA on a path leading to a universal capability to communicate with all national and international networks. In the transition to the ISO standards the cost of the hardware and software involved will be substantial, and NASA should begin planning for this expenditure in a few years.

Recommendation

CODMAC encourages the NSI coordination function and the development of applications software and network interfaces that can support interoperability and survive the transition from current protocols to the ISO/OSI suite.

Program Support Communications Network Utilization

The NASA Office of Space Operations (OSO Code T) Program Support Communication Network (PSCN) program offers a potential mechanism to meet the immediate and future communications needs of the NASA Office of Space Science and Applications (OSSA Code E) for its multidisciplinary science community. These communication needs are best served by a nationwide, full service network linking science users with one another and with flight projects, data archives, and computational facilities. This networking capability should address the immediate needs associated with science analysis and should evolve to support a teleoperational environment during the Space Station era. As noted previously, the science community of OSSA should be extensively involved in the design and implementation of a science network. The community should determine the functional services and protocol of the network, the user sites involved, the internet connections to other networks, the priorities of implementation, and system trades of functionality versus costs. Currently, the scope of the PSCN provides the communication backbone and tail circuits to potentially link NASA Centers and the science community but stops short of providing networking services over these links.

CODMAC applauds the formation of the PSCN Users Committee by NASA Code EI and TS and fully supports the implementation of the NASA Science Internet (NSI) as outlined in its Project Plan. The establishment of this committee and the development of the Project Plan by the committee directly follow the precepts of science involvement as described in CODMAC I and II. The proposed involvement of the PSCN Users Committee in the implementation and operations of the NSI is also supported.

It is recognized that the PSCN will be under programmatic and resource pressures during the development of the NSI. CODMAC believes that the Users Committee should serve as a mechanism to help make the trade-offs between implemented functionality and costs. The committee could be used to specifically support network topology and line consolidation studies to maintain the desired communications functionality within the cost constraints of PSCN.

It is now planned that Code T overall support for PSCN utilization, tail currents, etc., be halted. Instead, it is proposed that individual research groups (at universities, NASA laboratories, and non-NASA institutions) be funded by their sponsoring Code E organization. This proposal seems to be undesirable: as funds are disseminated to research groups, they will be taxed and/or burdened as are other research funds. This will reduce substantially the funds actually available to support networking connections and related activities. Moreover, the cessation of Code T support and coordination of NASA networking connections will effectively eliminate the effort to reduce the number of tail circuits. We conclude that only by centralization of oversight and funding authority can NASA networking links be made efficient, nonredundant, and cost-effective.

The formation of the NSI program and the PSCN Users Committee sets important new directions for NASA networks that support science. Within this structure it will be possible to consolidate the diverse physical networks and provide cost-effective and highly functional services to the broad science community. Codes E and T are to be commended for defining and supporting this activity. They are strongly encouraged to find ways to ensure that these vital services to the NASA centers, and especially the distributed user community, are adequately supported and funded.

Recommendation

CODMAC recommends that NASA ensure adequate support and funding for the NSI and related activities, including intercenter communication and user tail-circuits. CODMAC further recommends that the initiatives established in the PSCN Users Committee and the NSI project be continued to ensure adequate user involvement and review in the development and operational processes.

4
Supercomputing for NASA Funded Sciences: Resources and Access

SUMMARY

This chapter attempts to highlight a serious problem: the lack of supercomputer facilities for NASA-supported scientists to perform research in areas ranging from data reduction and analysis to sophisticated multidimensional numerical modeling of planetary magnetospheres. It is our feeling that this problem may considerably reduce NASA scientists' ability to effectively and efficiently utilize the resources that NASA will spend billions to produce as the Space Station era begins.

After examining supercomputer utilization and access within NASA, CODMAC has found that the computational needs of the NASA research community are not being met; there exists no centralized means of evaluation and utilization of computational resources; access to computational resources is inadequate especially for the scientists at a center without a supercomputer; and there has been little input from users outside of NASA centers on issues concerning procurement of supercomputers, the operation of NASA computing resources, and access to those resources. While it is possible that NASA has in toto adequate supercomputing resources, it is clear that the allocation of these resources for scientific users supported by OSSA is not sufficient.

The committee recommends the following for addressing these issues:

1. NASA should develop an agency-wide plan for the management of its supercomputer resources.

2. OSSA should prepare a strategic plan for supercomputing for NASA space science users.

3. The OSSA centers should recognize and anticipate the needs of their user communities and be an advocate for them when procuring resources for scientific computing.

4. Users of supercomputing should be involved at every level in the development of the supercomputing plans and should participate in the oversight of the operations of NASA's supercomputing resources.

INTRODUCTION

The evolution of computers in their many sizes and speeds has made a profound impact on the methodology of science. Computers are now pervasive in all aspects of science from the collection of data to the evaluation of the most advanced theories. Herein CODMAC conveys its concerns about NASA's utilization of a specific class of computer, the supercomputer, in the sciences that NASA supports. The term supercomputer is used to describe a large mainframe computer that operates on cycle times measured in nanoseconds. These machines take several forms: the common single processor variety, such as the CRAY 1, the multi-processor machines, such as the CRAY X/MP and CRAY 2, and the parallel processing machines that have thousands of processors working together to produce aggregate speeds at the present CRAY 1 machine level or above.

It is the feeling of CODMAC that there currently exists a significant need for supercomputer-class machines within NASA's science community that is not being met. This need ranges from computer modeling—to assist in understanding the data NASA has and is now collecting from space—to processing of spacecraft data. Further, there is a grave concern that the capabilities of the next generation of instruments to be flown on the Space Station, the Great Observatories, and planetary missions will produce data streams that will completely overwhelm the ability of any of the available supercomputers to process them, even at the crudest level. This latter problem is addressed in more detail from a different perspective in Chapter 5.

NASA has entered the supercomputer era and now possesses nine supercomputers at its various centers. They range from a VPS-32 at Langley to a CRAY 2 at Ames Research Center (ARC). One of the computers is a one-of-a-kind parallel processor machine, the Massively Parallel Processor (MPP) at GSFC. Of these supercomputers only three are generally available to NASA scientists and the scientists that NASA supports within OSSA. The three supercomputers are the CYBER 205 and the MPP at GSFC and a maximum of 10 percent for science users on the NASA CRAY 2 machine at ARC. However, the 10 percent available on the Numerical Aerodynamic Simulation (NAS) computer at ARC is to be split 90 percent for aeronautical and fluid dynamic research and 10 percent for research outside these areas. This leaves only "1 percent" for space science and other Code E users.

CODMAC has reviewed computational resources in meetings at GSFC, ARC, and Jet Propulsion Laboratory (JPL). Its findings are that the CYBER 205 at GSFC has been represented as fully utilized, the MPP is not a machine useful for the general mix of scientific problems on a production basis, and the ARC CRAY 2 has limited availability. JPL has conducted a study that shows that at least 2/3 of a CRAY 1-class machine is currently needed by the scientists affiliated with JPL. This requirement alone exceeds the identified resources available to Code E scientists.

In considering the present supercomputing environment within NASA, CODMAC feels that the topic of supercomputers and their utilization has not been adequately addressed. Yet these machines continue to be purchased. In the following paragraphs we have attempted to elucidate the salient issues as the committee perceives them and offer some recommendations for their resolution. The issues listed and discussed are by no means all that exist but CODMAC feels they typify the supercomputing situation at present.

ISSUES

The thrust of this chapter centers on the need for a coordinated NASA policy to provide adequate supercomputing computational support for the sciences within NASA and for NASA-sponsored researchers. This section of chapter 4 identifies four issues that ought to be addressed before any such policy can be designed or implemented.

The issues are summarized below, and a more detailed description follows.

Summary of Issues

- *The computational needs of the NASA research community are not being met.*

The supercomputer computational requirements of NASA and NASA-sponsored researchers are not being adequately met. The current need is being fueled by an ever-increasing pool of researchers who depend on supercomputers for an efficient and cost-effective means of data processing, modeling, simulation, and algorithm testbedding. Projected future computational requirements, which have not been adequately assessed on a NASA-wide basis, will dramatically increase the use for computer resources. Unless more supercomputing cycles are made available to the NASA scientific research community, data processing will be a significant limiting factor in the next generation of scientific results. For example the high bandwidth EOS data alone appear capable of completely overwhelming the current processing capability. A coordinated effort between NASA, OSSA, and the centers must be undertaken to effectively gauge the current and projected computational requirements of their respective research communities.

- *There exist no centralized means for evaluation and utilization of computational resources.*

At this point in time, no means exist within NASA or OSSA to ascertain the availability of computational resources. This situation has arisen from the fact that procurement of the NASA supercomputers has been accomplished largely through program or discipline-specific initiatives without a high degree of centralized coordination. Of the supercomputers in question, only the CYBER 205 at Goddard, a maximum of 10 percent of the CRAY 2 at ARC, and the MPP have been identified as potential resources for the NASA research community within OSSA [Notes 1 and 2]. It is not clear if other facilities, such as those at JSC and MSFC, are willing or able to provide support to this community. In reviewing the results of a study performed for Code E by

Science Applications International Corporation (SAIC) [Note 3], CODMAC noted that NASA computer centers did not respond to the written request for information while those at NSF centers, universities, DoE organizations, and NCAR responded. It is CODMAC's opinion that this lack of response represented two major problems within NASA's computer hierarchy: (1) the lack of responsiveness on the part of the various NASA computer centers, and (2) the inability of the contractor doing the study to identify an individual or organization within OSSA or NASA that possessed even rudimentary information concerning the various computer centers, exemplifying the situation that prompted this investigation [Note 4]. In order to address this situation, NASA and OSSA should establish single points of contact for their respective research communities from which relevant information on the availability of computational resources can be obtained.

- *Access to computational resources is inadequate.*

The notion of access to computational resources includes not only the concept of connectivity, but also the process of allocation, the level of service in user and software support, and cost. Currently, efficient, high-bandwidth access to the GSFC CYBER 205, the ARC CRAY 2, the MPP, and other NASA facilities is virtually nonexistent for the NASA research community that does not reside at one of these centers. Further, software support and user services functions appear to be inadequate. NASA should address this issue by expanding user interaction in network development and support requirement activities. It should be noted that NASA is in the process of creating the backbone of such a network with the PSCN and a NASA computer network, NCN, that is currently planned to connect Code R supercomputers.

- *User involvement in exploitation of computational resources has been inadequate.*

The involvement of the NASA research community in all phases of computational resource utilization has been inadequate. It is imperative that the whole research community, not just personnel at a specific center, be given the opportunity to participate in the discussions concerning computational issues. Specifically, the procurement of new hardware and new software, and the level of the facility services should be areas where the general scientific community has input. Such involvement, for example, would help ensure the smooth flow of large quantities of observed and model data that is missing from many present systems into and out of the supercomputer. Finally, the total community should be involved when the inevitable cost trade-offs must be made because such trade-offs have an impact on the research being performed in a significant manner.

Detailed Description of the Issues

Meeting the Needs of the NASA Research Community

The supercomputer has become a vital tool in both basic and applied research. In the areas of data analysis, modeling, simulation, and algorithm testbedding, researchers are allowed to address problems of a scope and complexity never before possible. The practicality, potential, and cost effectiveness of the use of supercomputing has been demonstrated in every discipline represented within Code E and in many multidisciplinary applications as well.

A good example of this is the CRAY 2 at ARC. It has the computational power to effectively replace the wind tunnel in aerodynamic research in a very cost-effective and timely manner. There seems little doubt that modeling in other scientific disciplines of Code E would or could yield increased scientific productivity at a reduced overall cost. Examples can be found in the space sciences where computer modeling could help define the questions that a given space instrument should be addressing thereby leading to improved and more effective sensor design and a better definition to the data acquisition, analysis, and storage requirements. This ability to assist mission planning and scientific analysis has, to date, been largely ignored. The DoE and DoD have been using supercomputers, the largest and fastest computers available, in this fashion with considerable success for several decades and continue to do so.

Currently, the NASA research community has potential access to the CYBER 205 at GSFC, and a maximum of 10% of the 4-processor CRAY 2 at ARC. Assuming for the sake of argument, that the CYBER 205 and each of the CRAY 2 processors is approximately equivalent to a CRAY 1, there would be approximately 12,000 CRAY-1 hours available to the NASA research science community within OSSA per year. Contrasting this number with current demand is difficult because no attempt has been made to quantify this demand on a NASA-wide basis. However, a JPL study for the supercomputing requirements of the planetary community alone has identified a need of over 5,500 CRAY 1 hours per year (2/3 of a CRAY 1 year). The saturation of the GSFC CYBER 205 yields 8760 CRAY 1 equivalent hours that are being utilized annually. Thus, on the basis of current use of the GSFC CYBER 205 and projected use by the planetary community, the demand for supercomputer cycles by the OSSA research community (14,260 hrs.) has already exceeded supply provided by the CYBER 205 and CRAY 2 [Note 5]. The addition of the requirements from the other disciplines within Code E will only widen the gap. Thus, it appears that NASA cannot currently meet the estimated need of its research community for supercomputer cycles. The MPP is not considered in this estimate because of its ineffectiveness in running production calculations and the start-up overhead required to begin using it.

Two aspects of future utilization will make the problem worse. First, the amount of data that will have to be processed, reduced, and analyzed will dramatically increase in the Space Station era with a projected data rate of 10^{12} bits/day. (It is interesting to note that the IPAC/IRAS data analysis team currently saturates a CDC 180/840 and an IBM 3032 as well as utilizing three image processing workstations.) Second, the number of researchers who will utilize supercomputers will continue to grow as the potential of supercomputers is documented in the literature and as more researchers break out of a "computational inertia" brought on by the lack of computational options available. The fact of the matter is that unless additional supercomputing cycles are made available to the NASA scientific research community, data processing will become the limiting factor in useful science production in the 1990s.

Centralized Evaluation and Utilization of Computational Resources

No mechanism currently exists within NASA or OSSA to obtain information on the availability of computational resources. Recently, GSFC and ARC have begun to make such information available, but in both instances, the announcements of availability were center initiatives (for example, see Note 2) with no coordination or centralized direction. An example of how difficult it is to gather the appropriate information can be found in the experience of a contractor, SAIC, conducting a study for NASA Headquarters. Not one NASA supercomputer facility answered a questionnaire sent to them, although cooperation had been requested in writing by Code E. The lack of general oversight at the OSSA or NASA headquarters level implies to CODMAC that the ability to coordinate the efficient use of supercomputer resources or provide general information to scientists under NASA's umbrella does not exist. The committee feels that the efficient allocation of present resources and any future procurements of computers of the supercomputer class requires oversight by NASA across codes and centers.

The inability of NASA to meet the demands of its research community has forced many researchers to purchase cycles from other sources (i.e., Los Alamos National Laboratory and NCAR). While this situation may be a necessity, it can result in a tremendous waste of effort if users must use several quite different systems to perform their research. Efficient utilization of each new facility requires proficiency in and knowledge of the relevant hardware/software/user support and access parameters. In addition, by purchasing cycles from an outside source, NASA gives up some control of costs. Nevertheless, certain scientific disciplines within OSSA find this method of supplying the necessary computational resources preferable to dealing with the limited resources available within NASA.

NASA currently has a distributed supercomputer network that includes two CYBER 205s, one VPS-32, one CRAY 1, one CRAY 2, three CRAY X/MPs, and an MPP. These machines use a variety of operating systems, software support, and user services functions (Table 1). PSCN provides a potential common access to all of these. A planned distribution system may well have contained many of these same attributes. However, with the exception of the CYBER 205 at Goddard and the 10 percent of the CRAY 2 at Ames, it is not clear to what extent this distributed system could be utilized to support the requirements of the NASA-wide research community. Are the other facilities willing or able to provide the level of support necessary to balance the supply and demand? If not, then other options such as purchasing

TABLE 1 Supercomputer Operating System

	Computers					
Centers	CYBER 205	CRAY 1	CRAY X/MP	CRAY 2	MPP	VPS-32
Goddard[1,2]	VSOS				Special Purpose	
Langley[3]						VSOS
Ames[4]	VSOS	COS	COS	UNIX V		
Marshall[5]			COS			
Lewis[6]			COS			

[1] The CYBER 205 must be accessed through an IBM 3081 and one of two AMDAHL machines (V6 or V7) operating systems that are like IBM.

[2] MPP is special purpose but users must be knowledgeable in VAX VMS Operating System, as well as MPP's own system, which is unique.

[3] VPS-32 is CYBER 203 upgraded to architectually resemble a CYBER 205 front-ended with a CYBER 170-370 operating system NOS.

[4] CRAY XMP/48 is changing operating systems from COS to UNICOS, a UNIX type operating system.

[5] Using an IBM 3033 as a front-end machine probably uses MVS operating system.

[6] Marshall machine front-ended by IBM 3084 using MVS operating system.

cycles, leasing new machines, or purchasing new machines should be considered.

Access to Computational Resources

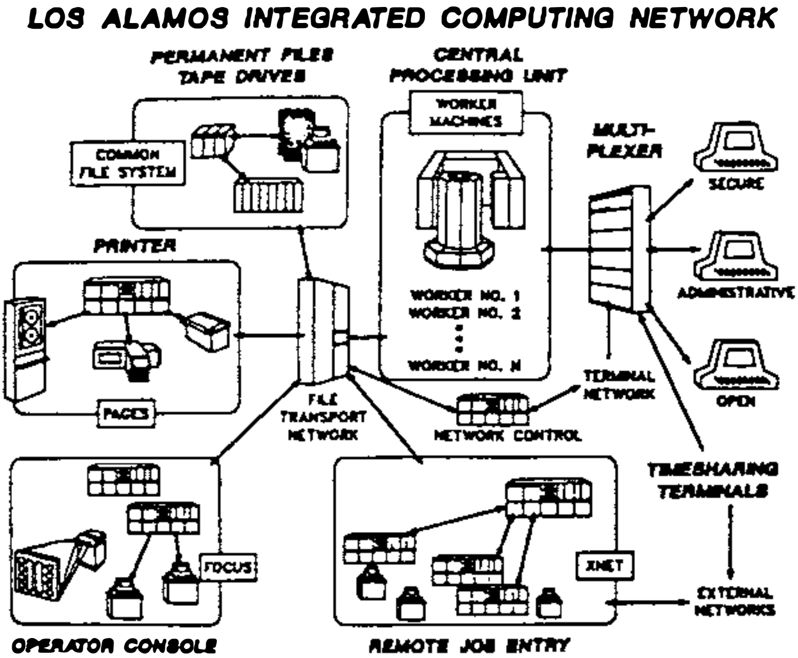
The NASA research community will place several requirements on any supercomputer facility that it may utilize. First, efficient and high bandwidth access should be provided to the computer facility for both internal (i.e., at a NASA center) and external NASA researchers. In addition, users should expect to interface with only one operating system per facility. There should also be methods of assessing job status while it is executing. Second, an allocation and priority procedure ought to be established and coordinated on a NASA-wide basis. Third, NASA should provide comprehensive support in areas such as software support and user service functions. Particular components of this may be centralized or distributed as needed.

For example, a suitable center has the central worker CPUs, but it must also have the proper front-end and back-end components. Figure 1 shows the configuration of the computer center at Los Alamos National Laboratory. At the front, one needs good terminal access systems; at the back of the installation one needs large mass storage, and printing and graphics facilities. Further, one needs optimized production operational control to level the loading between computers. Either at the front or back, one needs external network connections. Some NASA centers have such a system, most notably ARC, but others cannot, are not, or will not develop the capabilities and services that are needed to make a complete computer facility.

User Involvement

As stated earlier, the level of user involvement is crucial if NASA expands its current computational facilities or procures new ones such as the new supercomputer to be purchased for JSC. Many important questions must be asked before appropriate trade-offs and decisions can be made. Are the users willing to compute at a remote site if the facilities are or could be made superior? As an example, NASA ARC currently operates a variety of supercomputers and a substantial number of support facilities. Is it more cost effective to place other supercomputers at ARC and have that center run the computer for a given discipline than have that discipline or center procure its own computer and support facilities? Would users be willing to use time at a variety of centers that possess supercomputers if fragments of these machines are made available for research external to that center? What software do the users in various disciplines consider essential, very useful, or of little use? Have the scientists in the solar-terrestrial theory program been satisfied with the support provided at Los Alamos? Should that support be expanded to other science disciplines thus avoiding the need for NASA to purchase additional computer power?

The committee feels all of these questions and many more will help enable NASA and OSSA to provide adequate resources for the science it supports. Further, the involvement of the users



- Worker computers (computers that execute user code)
- A Common File Storage system (CFS)
- A Print and Graphics Express Station (PAGES)
- A Facility for Operations Control and Use (FOCUS)
- The Extended Network Access System (XNET)
- A terminal network with both physical and logical security partitions

FIGURE 1 Los Alamos integrated computing network.

may lead to a much more cost-effective use of the supercomputer environment than currently exists within the NASA structure.

Recommendations

1. NASA should develop an agency-wide plan for the management of its supercomputers.

The problems in the management of NASA supercomputers extend across the NASA divisions. We have identified problems in Code E that we believe can only be solved by an agency wide approach. As noted, NASA has an impressive array of supercomputers that form a major national resource. Yet, many NASA science users have serious problems obtaining information about NASA computers and access to them. This point was given emphasis by a recent survey of scientific supercomputing capability that was funded by Code E. Code E requested information from all of the NASA supercomputing facilities as well as those operated by the Department of Energy (DoE) and the National Science Foundation (NSF). Both DoE (the Los Alamos National Laboratory and Magnetic Fusion Energy Computer Center at Livermore) and the NSF (NCAR computing center and members of the NSF Supercomputing initiative) cooperated in this survey as did several universities. As related above, none of the NASA supercomputing centers responded. The barriers to the use of computers by users funded by a code other than the one that owns the computer are substantial. Until very recently there was no mechanism by which users funded by Code E could gain access to machines purchased by Code R. Only recently has a very small fraction of the time on the Numerical Aerodynamic Simulation (NAS) been available to users who are not funded by Code R [Note 2].

CODMAC believes that the issues outlined in this chapter can only be resolved if there is coordinated supercomputer management within all of NASA. Only with coordinated management can users be assured of access to a computer system suitable for their needs. In developing the plan for coordinated management, it is suggested that NASA review the management schemes used by other agencies such as the Department of Energy and the National Science Foundation.

2. OSSA should prepare a strategic plan for supercomputing for NASA space science users.

Within OSSA there is no plan to provide supercomputing resources to space science users. Such a plan is needed. While our investigations show considerable need they do not represent a complete survey of the needs of the entire space science community. We recommend that OSSA undertake such a study as the first phase in developing this plan. So far, only the planetary community through JPL has done this. This survey should try to estimate future growth in computing needs. We believe that the growth will be significant. Disciplines such as solar-terrestrial physics, in which there is considerable experience in using supercomputers for research, have experienced an exponential growth in resource needs. OSSA should carefully evaluate the types of resources needed by its community. They should evaluate the needs for large scalar and vector processing and the needs for parallel processing. They should also evaluate the needs for special purpose computing to support specific missions or experiments. However, we recommend that this be done with extreme care. NASA originally developed the Massively Parallel Processor (MPP) to support the LANDSAT missions (see Note 6). When LANDSAT processing was moved away from NASA, NASA was left with a special purpose processor without a purpose.

Once OSSA has determined the user needs, the office should formulate a plan to meet those needs. However, this does not necessarily mean that OSSA should plan to buy new supercomputers. They should also consider using resources from other NASA codes and resources run by other agencies that have more experience in managing supercomputers than does OSSA.

3. The OSSA centers should recognize the needs of their user communities and be an advocate for them when procuring resources for scientific computing.

The OSSA centers (GSFC and JPL) have the closest contact with the scientific community. They are frequently the institutions that actually provide services for scientists. They have a scientific constituency that extends beyond the center gates and includes many scientific members of missions operated by the centers. Unfortunately, centers have not fully recognized this external constituency, and have not included them in their computer resource planning. The committee recommends that the centers include external users needs in the plans for supercomputing.

4. Users of supercomputers should be involved at every level in the development of supercomputing plans and should participate in the oversight of the operations of NASA's supercomputing resources.

Scientific users have a major stake in the development of the supercomputing resources. As noted in this chapter, in many cases, important research simply cannot be accomplished without supercomputing resources. Users also provide an indispensable source of the information necessary for NASA to make the proper choices of hardware and software. For instance, there are considerable differences between various supercomputers: some offer excellent performance when using large vectors but poor scalar performance, while others offer good performance for scalars and short vectors but poorer performance for large vectors. Which type to buy can only be determined by consultation with the user community. The committee recommends that users be involved in all stages of system design and implementation. It is also recommended that users be included in the oversight of NASA's computing system operations. This is the best way to ensure that the computing organizations remain responsive to user needs.

NOTES

1. Computational Physics Investigations Utilizing the Numerical Aerodynamic Simulation Processing System: Office of Aeronautics and Space Technology Program Notice, May 13, 1986.
2. Dear Colleague Letter for usage of Cyber 205 at Goddard Space Flight Center, June 24, 1986.
3. Earth and Environmental Sciences in the 1980's: Part I—Environmental Data Systems, Supercomputers, and Networks, Science Applications International Corporation. NASA Contractor Report 4029, NASA contract NASW-36Z2, October 1986.
4. Subsequent to the CODMAC meeting addressing the question of supercomputer usage and Mr. Hood's presentation of the results of his survey, the NAS at Ames, and NSESCC and MPP at Goddard and Langley responded. The Marshall group

did not. Lewis' stated response would not be forthcoming because information was already part of "the public record."

5. This assessment assumed saturation of the Cyber 205 and the availability of the full 10 percent of the NAS Cray 2. If the announcement specifications (2) stand, only 1 percent of the NAS Cray 2 will be available, and the shortfall may be significantly more than estimated.
6. Data Management and Computation, Vol. I: Issues and Recommendations; Committee on Data Management and Computation, National Academy Press, Washington, D.C., 1982.

5
The Management of High Data Rates and High Data Volumes

SUMMARY

Sensors and communication systems with high data rate and volume capacities are being planned for the Space Station era, which will present challenges for efficient acquisition, processing, and management of data. It is essential that a strategy be developed to manage these expected data rates and volumes with the goal of providing the greatest scientific return within a set of resources. The data acquisition strategy should be driven by the scientific requirements taking account of the data system resource availability and life-cycle costs. In supporting this strategy, scientific users must have access to appropriate tools and techniques to plan and execute the observing sequences. Meeting these challenges will also require an aggressive technology development program for high-rate/high-volume data and information systems. Management of data rates is part of a larger complex of issues that will arise as a result of the eventual availability of suites of sensors with flexible observing characteristics.

The committee's recommendations for addressing these issues are as follows:

1. The goal of any data management strategy must be to maximize the scientific return from the acquired data within the constraints of the data system. Users must be given an active role in

the planning of data collection both to ensure high scientific return from the data and to increase awareness in the users of the implications of various observing scenarios.

2. Future space science missions should develop a data acquisition strategy that allows the user to participate in the data management process. This participation should include consideration of total life-cycle costs, interactive payload control to allow data editing based on quality, and mechanisms for conflict resolution.

3. OSSA must aggressively pursue the development of tools and techniques that will enable a robust data rate management strategy to be adopted. Performance models should be developed. Testbeds of methods for on-board information extraction and autonomous instrument control should be implemented and data compression techniques should continue to be pursued.

BACKGROUND

Future space and earth science missions will have much larger data-gathering capabilities than previous or present missions. There are several contributing factors to this overall data-rate management problem. First, there are limitations in the potential data transmission rates between satellites and the ground. This limitation is more acute in planetary missions because antenna size and power availability restrict the potential transmission rate. For earth science missions, the availability of the Tracking and Data Relay Satellite System (TDRSS) link should provide an adequate data transmission rate. However, some of the Earth Observing System (EOS) high-rate sensors are near or beyond the TDRSS limit, and possible conflicts with other contemporaneous missions or with operational and commercial sensors aboard EOS may further reduce the capability of the TDRSS link to serve science users. Second, our ability to provide long-term data storage is limited relative to our ability to collect data. Although data storage technology continues to improve, the use of high-data rate sensors and the increased scientific use of long-time series data from multiple sensors have greatly increased the data storage requirements. Third, the use of several data types from multiple sensors to study certain processes will require sophisticated data handling techniques to cope with the resulting high data rates and complex data stream. The command and control of these varying sensor ensembles are much larger problems, of which data rate management is a major component. Fourth, networks will be required both for transmitting commands and observing scenarios and for data transmission. The potential data rates over these networks are large and will require sophisticated planning in a heterogeneous computer environment.

A variety of technology issues arises in data rate management. Although there is a considerable body of knowledge in data volume reduction, there remain several areas that need additional development. In the area of data compression, which maximizes information carried on the available bandwidth, there are questions concerning noiseless source coding and application-dependent and instrument-dependent compression. In the area of data editing, there are questions concerning autonomous editing (as opposed to static or deterministic editing) and evaluation of its performance. There are also technical issues in the management of the observational and data system resources, particularly in the design of observing scenarios that can be evaluated in terms of their data requirements.

The solution of these technical problems must be developed closely with the science users as the potential solutions may greatly affect the scientific quality of the data. Any method of data volume reduction must be designed not to deteriorate the scientific return. As various observing scenarios may have conflicting needs or exceed the capabilities of the observation and data system, there must be a management strategy for resolving these conflicts.

CODMAC emphasizes that data rate management goes beyond the technical issues of data compression and editing. They are merely some of the techniques that can be used in the overall tool of data-rate management. Management does not necessarily imply volume reduction. Rather, it implies the coordinated use of a variety of techniques to cope with high data rates and large volumes in a sensible way such that the scientific value of the data is not compromised.

DATA COMPRESSION AND DATA EDITING— THE TECHNIQUES

Two fundamental techniques for data rate and volume reduction are data compression and data editing. Both are useful for reducing data transmission and storage requirements. Data compression, the most typical approach, involves the elimination of redundancy in the data. However, substantial overhead can be incurred in the encoding and decoding process. Data editing can be used when there is reason to believe that the cost of acquiring some segment of data generated by an instrument is greater than its value for a particular application. It is perhaps the most effective way of controlling data rates and volumes because it simply eliminates much of the marginally useful data right at the source.

Data Compression

Data compression typically involves the establishment of a base value for some contiguous block of data, with subsequent values in a time or spatial series encoded relative to the base value. Two-dimensional data, such as image, are generally treated as a single time series. As dimensionality increases, compression can be increased further by coding variances in more than one dimension. Other common compression schemes involve transforms to the frequency domain, classification of pixels according to some similarity measure, and fitting the data series to some mathematical model that can be reconstructed parametrically. The elimination of redundancy, however, makes the data stream more sensitive to bit error rates and thus can have the effect of increasing link performance requirements.

All distortion-free compression algorithms are constrained by Shannon's noiseless coding theorem to rates greater than or equal to the source entropy. These distortion-free compression algorithms may be generally classed as "fixed rate" or "variable rate" source coding, and all are prefaced by various techniques for removing data redundancies to transform the actual source into one that approximates a memoryless source. As a specific example, we may take a serial, sampled data stream with a high degree of first-order correlation and transform it to a nearly uncorrelated data stream by performing successive difference operations. To implement a variable rate source code, we might then assign the number of bits to represent a certain data value based on the probability of occurrence of that data value for the particular source (this is usually called Huffman coding). One problem with this approach is that the source statistics are in general not stationary, and must be included in the processing memory and along with data to maintain running estimates of the source probability density function that is used to make bit assignments.

If a data compression algorithm introduces no distortion (i.e., the input data can be fully reconstructed), and throughput performance is adequate, then there is no real implication for the user. Distortion-free data compression can often reduce data rates by factors of 2:1 to 3:1. The theoretical limit for many image data types (which typically generate the most data) is probably closer to 6:1. However, if the original data are not fully recoverable, the user will be concerned with the acceptability of the level and type of distortion.

When considering very high data-rate instruments (greater than 10 Mbps), the performance of the data compressor hardware and software becomes an issue. Current flight implementations are capable of handling 1 Mbps rates from instrumentation. Clearly, this technology needs to advance significantly in order to be employed in missions being planned for the mid-1990s, which may carry instruments capable of peak data rates on the order of 300 Mbps.

Data Editing

Data editing approaches to data-rate and data-volume reduction may be deterministic or autonomous. In either case, the potential for limiting data acquisition is application dependent and must be user controlled. Any substantial amount of data editing is also predicated on the ability to optimize for a single purpose. If data are to be shared among several investigators with conflicting requirements, some form of arbitration then becomes necessary. Only by making the user requirements known in the beginning and by using those requirements to drive the data acquisition process, does the data editing approach become a viable technique.

In a deterministic approach, the user knows a priori what portion of a data stream to acquire, and it is a straightforward job for the on-board data system to filter out the undesired data. The experimenter determines data needs in advance of a data acquisition opportunity and requests only the data of interest, controlling such parameters as coverage, frequencies, channels, look angles, and so on.

In an autonomous approach, an on-board processor applies an information extraction algorithm to the data, the results of which are used to determine the desirability and quality of the acquired data. If the data do not satisfy some quality criteria, which itself can be dynamic, then the data are not transmitted. This more automated approach to limiting data acquisition at the source is potentially a very effective means of data-rate and data-volume reduction—particularly for systematic repeat-coverage scenarios supporting long-term time series—but will require significant technological advances in on-board computing capacity, very careful selection and validation of algorithms, and the development of mechanisms for user interaction and arbitration.

RESOURCE MANAGEMENT—THE TOOL

Recognizing that any data and communications system is likely to represent a constraint on the data acquisition potential of an instrument or suite of instruments, it is important that the data generation and handling resources be managed in such a way as to maximize scientific utility. Therefore, the inclusion of operational modes and support tools into the operations system that allows the investigator to accomplish the management function by communicating requirements, should be considered a necessary element of any strategy to deal with high data-rate and volume instruments. In a complex data system with dynamic mission scenarios, it is the investigator who must ultimately make the trade-offs in planning the use of limited resources. Also, as many earth science data sets will be used as part of long-time series, the science usage is not limited to the individual who has immediate interest in the data.

Many scientific spacecraft of the future will be highly complex systems. There will be a multitude of instruments supporting single-instrument, multi-instrument, and multidisciplinary investigations. Many of these instruments will be gathering data continuously and others will have highly variable duty cycles. Data system resources, such as computational and transmission bandwidth, on-board storage, power, and ground storage will, as in the past, continue to be fixed, or at least bounded. As more is known about the instrument behavior and the potential data applications, the scientific scenarios will change. The challenge is to apply the fixed resources in the most effective manner possible to support highly dynamic usage. This becomes particularly crucial for very high data-rate instruments because they are such major consumers of resources.

Networking for transmission of data sets and command sequences will continue to be important in the overall data system. Although potential data volumes will likely outstrip the network's capability for transmission, the same procedures for data-rate management can be used to increase network performance. Sophisticated data compression/decompression methods and automated data transmission procedures can be employed so that the system load is more evenly spread throughout the day. The additional challenge in this area will be the presence of a heterogeneous computer system. Data management routines must be able to work on a variety of computers, rather than being limited to a few well-defined systems present in the data observation and acquisition system.

A resource management strategy in support of a data-rate management function can be enabled through the planning of requirements-driven experiments. That is, rather than the investigator taking all the data possible at a particular opportunity via explicit instrument control, the data acquisition process is interactively optimized based on parameters of coverage, time, quality, quantity, and data type. This will require sophisticated planning tools, autonomous instruments, and the development of much more useful notions of data quality. The planning tools must fit within the overall command and control structure for the spacecraft operations system. The goal is user allocation and control of observational and data system resources, allowing the user to make the trade-offs necessary to maximize the data acquisition within resource bounds for a particular application.

Recommendations

1. The goal of any data management strategy must be to maximize the scientific return from the acquired data within the constraints of the data system.

It is essential that users become involved in the planning of data collection strategies, in part because of the flexibility of future observing systems and in part because of the potentially high data rates from future sensors. The design of data-collection and data-management strategies involves decisions that can greatly affect the scientific quality and usefulness of the data. Our overall recommendation is that users be given an active role in planning and controlling data collection, not only to ensure high scientific quality in the data, but also to increase awareness among the users of the implications of various observing scenarios on data-rate management. The planning of data-collection strategies must also include a mechanism for arbitrating the inevitable conflicts that will arise between potential users of the observing system. This arbitration should aim towards accomodating all possible users within the constraints of the data collection and management system while not sacrificing the scientific quality of the data.

2. Adoption of a requirements-driven strategy for data acquisition, supported by appropriate planning and operations tools, allows the user to participate in the data management process.

We recognize that the flexibility and potentially high data rates of future observing systems will severely tax existing data-management systems. Decisions regarding data acquisition and storage will require that the interested science users be closely involved in the data management process since these decisions will greatly affect the scientific quality of the data. The availability of tools for planning and controlling observing scenarios will increase the amount of user involvement and user awareness in the management of data. The following specific steps should be taken.

a. Integrate the concept of life-cycle cost into data acquisition planning. Each data acquisition session should not only consider impact on mission data system resources but also on science data processing and management resources.

b. Provide interactive payload control. In the absence of adequate autonomous methods, such control will enable first order data editing based on data quality and value.

c. This strategy must include mechanisms for resolving potential conflicts between various observing scenarios. The presence of adequate planning tools should allow scenarios that accommodate all users to be identified.

3. The development of tools and techniques will enable a robust data-rate management strategy to be adopted and should be aggressively pursued.

We recognize that a certain amount of technology development must take place before we can realize the goal of close user involvement in data management. In the past, observing scenarios have been relatively fixed at launch; future systems will be much more flexible and will involve suites of instruments, rather than single instruments. Systems that will allow comparisons between observing scenarios must be implemented. Studies of advanced data compression techniques must also continue. We also note that the data rate management of these observing systems must be a marriage of technological tools and science requirements. The following specific steps should be taken.

a. Develop mission-specific, end-to-end data-system performance models that can allow thorough user analysis of trade-offs between data quality, and quantity, and resource usage. Modest changes in science requirements for extremely high-rate instruments can result in major differences in the data system design accommodation. Performance models would allow earlier insight into the technological requirements for instrument and spacecraft development and would provide a mechanism conducive to scientific participation in the design process.

b. Begin an active program, utilizing a testbed approach, that integrates data-rate management into an overall command and control strategy for instruments that are likely to stress available resources. This testbed should be a primary vehicle for developing and demonstrating methods of on-board information extraction and autonomous instrument control technologies.

c. Continue the development of data compression techniques and supporting data systems. Higher performance, noiseless source coding approaches to data compression are potentially available but must be based on careful analyses of the informational content of various data types. Research should continue in developing and understanding application-dependent data-compression techniques for instruments that are limited in some way because of high data rate or data volume. Research should be initiated that seeks to measure the information content of remotely sensed data with the intent of using the resultant metric as a means for controlling the use of data system resources.

CONCLUDING REMARKS

The issue of data-rate management will become increasingly important in future earth and space science missions. It is further complicated because it is a combination of technical and management problems. The fundamental goal is to maximize the scientific return from the mission within operational constraints. This includes both the spaceborne and the ground-based components of the data processing and storage system. The conflict between the scientific desire to acquire more data and the physical limitations of the data system needs to be addressed on two fronts. First, the science user needs to become closely involved in the planning of data acquisition scenarios. Although this requires the development of appropriate planning tools, the primary requirement will be the development of a management structure that will allow user involvement in the design of observing scenarios, increase awareness of the data system implications of particular observing scenarios, and resolve potential conflicts between observing scenarios. Second, there must be a program of technology development to reduce the data volume in order to effectively increase the data system capabilities. Such data compression techniques must operate within the constraint that the scientific quality of the data not be reduced.

Data-rate management is one issue that arises when ensembles of high-rate, flexible sensors become available in future missions. This new flexibility will require new approaches in system management. A program that closely involves the science user in the design of these observing scenarios will be essential for their success.

6
NASA Data Management Issues

SUMMARY

The Committee on Data Management and Computation has reviewed the status of data management efforts in the NASA Space Sciences. It finds that while there has been progress in this area, more emphasis is required if NASA is to meet its needs in the 1990s. CODMAC recommends that NASA adopt and implement an explicit data management plan for all space flight investigations. Further, NASA should provide sufficient resources for data archiving and guidelines for its implementation, and enforce the requirements that projects and principal investigators (PIs) properly archive and document their data. Through the NSSDC, NASA should develop procedures for the protection of the data archive from deterioration of media, hardware failures, and tampering by individuals. We reiterate the need for an active, distributed archive, managed in scientific data management units by each discipline in coordination with the NSSDC. Through the NSSDC, NASA should also develop easily accessible, standard catalogs and directories to its data archive, including to its distributed archives. NASA should continue to assess storage media and develop guidelines for its use in archiving. Through the Information Systems Office of OSSA, NASA should establish an advisory committee on data retention and preservation and associated concerns

and should establish ties with other agencies and the user community regarding the dissemination of and access to archived data. Finally and most importantly, NASA should support and promote the use of its data archives.

INTRODUCTION

CODMAC and researchers and engineers within and external to NASA have long been concerned with the problems of collecting, managing, and accessing space-acquired scientific data. While activities connected with the missions themselves attract the most interest and actions, the seemingly more mundane issues of handling, managing, and disseminating the scientific data being collected are given less than adequate attention. As the accumulation of data continues, however, the task of data management grows more difficult. We see growing problems with the loss of data and the difficulty of providing data access. NASA management must move quickly and firmly to adopt, support, and enforce an explicit data management policy and plan for all space flight investigations.

Both previous CODMAC reports addressed the broad problems of data management and a cohesive, NASA-wide approach. CODMAC finds, however, that while there has been movement toward addressing the problems, more emphasis is needed in the near term if NASA is to meet the requirements of the 1990s without major difficulties.

A number of recent NASA activities have demonstrated advances in this area: the initiation of the IRAS Processing Analysis Center (IPAC); the Land, Climate and Ocean Data Systems; the Planetary Data System; and the Space Telescope Institute data systems are excellent beginning efforts. The NSSDC has begun a program to address several aspects of the data management problem including an understanding between the Planetary Data System program at JPL and NSSDC that covers their joint responsibilities for data, directory/catalog, and archive activities.

A coordinated approach to NASA-wide data management involves not only a number of technological issues but also clear direction from management. OSSA activity needs to be more focused, with long-term goals and objectives clearly stated. The amount of data being collected continues to grow; the number of possible users/researchers is increasing; and the capability for new, more comprehensive research is available. A strong movement for a comprehensive program is needed now.

ISSUES

Previous CODMAC reports identified a number of issues for a comprehensive data management program including a clear management directive and set of objectives related to data handling, and a firm commitment for comprehensive user support.

Further, since data are now collected from a wide variety of instruments for many different disciplines and applications, a distributed archive is probably the best approach. Data sets relating to a specific mission or knowledge area would best be stored at the center or centers where the technical specialists reside who use the data most and know the most about it. This approach is particularly important during the active data collection and analysis period. Long-term archiving could be handled by a central organization (NSSDC), or arrangements could be made for deep archiving at the distribution centers. For such a distributed system, good directories and catalogs of the data are essential: our view is that an overall single, national directory be available as a point of access. This central point of access would direct the user to more detailed catalogs that would be maintained at the various archive centers.

In addition, a number of problems concerning collection, storage, documentation, and preservation of data need to be addressed. As a progression from previous CODMAC reports, this chapter provides some detail on various (but not all) issues that need to be addressed. The topics and subtopics discussed are not intended to cover all concerns or to be in any way exhaustive, but are intended to call attention to pertinent and near-term needs.

Organization and Management

To date, there has been insufficient movement toward a comprehensive NASA-wide policy for data management. As the acquisition of great volumes of diverse data continues to increase, the problem of the lack of an overall policy becomes more acute.

CODMAC has previously suggested that NSSDC, with advice and support from the Information Systems Office (ISO), provide for NASA's central policy-planning and implementation with respect to data. Currently, it appears to be the only active function where efforts could encompass the wider tasks being suggested. Performance in response to CODMAC's recommendations has been spotty. It appears that NSSDC is not consistently involved in end-to-end mission data-management planning and implementation.

Every mission/project should recognize the data management problem and accept participation on the part of an NSSDC (or ISO) person in the project planning and implementation phase. NSSDC must actively plan for and provide various support functions. Storage and the long-term or "deep" archive management should be provided for by NSSDC. In the case of mission repositories or databases, the project should provide for the handling of the data with advice, standard methods, and other guidelines coming from NSSDC. Overall data cycle (end-to-end) planning should be provided by NSSDC or ISO, which should include the migration of needed scientific and ancillary data to the archives, conversion as necessary, and, at least, citing in the directories.

A NASA Management Instruction (NMI 8030.3A) identifies project scientists as responsible for assuming flow of project data to an appropriate archive. This has been largely ineffective because many project scientists do not have a firm commitment to archiving, and believe that funds used for archiving purposes might otherwise yield more scientific results from PI research and analysis. The approach of placing data archiving requirements in Announcements of Opportunity and of appointing interdisciplinary scientists for data management and archiving on flight projects is encouraged.

Recommendations

1. NASA should have an explicit data management policy and plan for all space flight investigations; should emphasize the roles of NSSDC and ISO; and should encourage the acceptance of the plan by the missions/projects.

2. Division directors and PIs should understand, accept, and implement data management practices.

3. A project scientist/PI's statement of responsibility should include the data management requirement. Continued funding should depend in part on successful data management activities.

Data Acquisition

This area involves the handling of initial, active mission repositories, predirectory processing (initial application of indexing or identifying detail), the addition of processed or ancillary data as needed, structuring the active databases, the access to the mission repository and related active databases by the concerned scientists, the provision of data set copies as required by researchers, and the use of accepted standard forms.

Much important scientific data, acquired by currently and recently active NASA spacecraft, is not readily accessible to the general scientific community. One reason for this is the typically low level of concern among the members of a spacecraft's science team (whose data are generally accessible to one another on a collaborative basis) for making data available to the wider scientific community.

PIs are funded by project offices annually for data reduction, analysis, and handling. While there are general contractual obligations to deposit appropriate data in an archive (NSSDC), funds rarely come specifically earmarked for this purpose. Thus, PIs find themselves in a position of having to satisfy a general NSSDC-data-submission requirement with funds that could otherwise be used for scientific analysis. Given the apparent absence of penalties for noncompliance with the data submission requirements, many PIs disregard, or only minimally satisfy, the requirement.

The comprehensive interpretation of space science data often requires the existence of ancillary information such as trajectory data, laboratory spectra, or ground-based observations. Sometimes these data are acquired as part of a particular space mission but at other times are acquired separately. Often times while plans are developed for archiving the primary database from a mission, no plans are developed for the archiving of these ancillary data. We urge that NASA and NSSDC establish guidelines for archiving the ancillary as well as primary databases and include observing logs and similar information as appropriate in their directories and catalogs.

Recommendations

1. NSSDC should prepare a general policy and methodology as guidelines for data acquisition processes with due consideration for the fact that each program will be different.

2. As requested and as possible, NSSDC should participate in preparing guidelines for access to and search of active databases and mission repositories. NSSDC should prepare guidelines for the inclusion of ancillary data in their archives and include such items as logs of relevant ground-based observations in their directories and catalogs.

3. Project data management/archiving funds should be separately identified and protected from use as data analysis funds; project data management plans should identify the relative levels of data analysis and data management/archiving funds.

4. Projects should have a deputy project scientist (or equivalent) for data management. The individual should be knowledgeable in both science and data management, and should have a strong personal commitment to making the project data accessible to the entire scientific community in appropriate form and quality. This individual could be an employee of or responsible to NSSDC.

5. Overall funding of a PI for research on active databases should take cognizance of his or her compliance with the archiving and documentation requirements and policy. Perhaps selection for future flight opportunities should be contingent on appropriate data-related performance on past missions.

Distributed Archives

The concept of distributed data archives now appears to be widely accepted. Although the long-term archive may well be at NSSDC, subject matter specialists or certain NASA centers with specific need or expertise may store data at their locations for all users (the mission repositories or active databases). It should be understood that the individual storage locations will maintain the proper detailed catalog of their stored data with the proper links to the main directory, support and provide needed access to their stored data, and maintain it and update it as required. These functions also require considerations of security and database backup or degree of permanence.

Recommendations

1. NSSDC, with the advice and support of ISO and in conjunction with the various NASA archive and science R&D centers both in and out of NASA, should begin the planning of policies, guidelines, and requirements for use by the various centers that *might become archival centers. Maintenance and customer services needs must be included in these plans.*

2. Because of its importance, and considering the current technical climate (computer hackers) and political climate (terrorists), policies, guidelines, requirements, and methods concerning security and backup must be addressed by proper specialists with guidance from NSSDC.

Directories, Catalogs, and Documentation

The concept of a single central, high-level, on-line directory of acquired and developed data has been recommended often and now appears to be accepted. NSSDC has made a beginning in the task of developing such a directory, but continued effort must be encouraged. NSSDC should confer with various groups, including other pertinent federal agencies, as to the form and means of access to the directory.

Only a limited beginning appears to have been made toward the development of detailed catalogs of the data that would be the second level in a hierarchy of directories and catalogs. The directory would lead users to the appropriate, distributed, detailed, on-line catalog. It should be noted that each of the various pilot systems provides a form of a detailed catalog, but it is not clear that any one of them should be the final accepted form.

NSSDC (with advice from ISO) should take the lead in developing a list of needed documentation, forms, and standards. This should be done in conjunction with mission/discipline personnel.

Recommendations

1. NSSDC, consulting with the various scientific users, missions, programs, and other agencies, must develop general standards and formats for the directory and catalogs, and also well-publicized, easy access methods.

2. Required documentation should include instrument functioning detail, data standards and formats, system user manuals, and software listings.

3. A directory or catalog is of no value unless it is used. The development and promulgation of use of these tools must receive early and strong attention.

Storage Media and Technology

The physical means of storing the data is extremely important because of the increasing volumes involved. Low-cost, dense, easily handled and easily and widely used media and degrees of permanence of data are the criteria. A continuing study and assessment of the technology must be pursued. Although some efforts are already in progress at NSSDC and JPL, broader attention to the problem must be emphasized. It should be understood that different media may well be used at different stages of the storage/archiving process. Magnetic tape and disks, and optical disks (both CD-ROM, and write-once-read-many-times types) appear to be the current technologies to be considered.

Recommendations

1. Pursue a plan for continued assessment of storage media and technologies with provision for distribution of the information.

2. NSSDC, in conjunction with mission/discipline representatives, should begin to develop guidelines for the type of media to be used at various stages in the storage process with due consideration for the overall data-cycle requirements.

Retention, Preservation, and Access

Data and information are useless unless they can be accessed in a relatively easy fashion; and on the other hand, too much data of varying degrees of quality may be unmanageable. Therefore, consideration must be given to developing guidelines for periods of data retention and decisions for purging.

Other problems of concern involve error in writing to media and failure of the media over time; conversion of digital data to other digital forms and to other kinds of media (print, microfiche, etc.); the provision of relatively easy, well-publicized, and speedy methods of access at any level in the data cycle; and the recognition that because of federal government regulatory and budget requirements, a policy of charges for access by non-NASA organizations may be required. Planning for these things cannot be done in a vacuum: users, scientists, and other agencies will be affected, will have serious specific concerns, and will need to be consulted. There would also be the exchange of knowledge about what other pertinent data sets exist outside of NASA. This could perhaps be accomplished through interagency coordinating committees. These groups should also have input for the problem of charging for data if it becomes necessary, a comprehensive marketing/publicizing plan, and a broad dissemination policy.

Recommendations

1. ISO should establish an advisory committee on data retention and preservation and associated concerns. Representation should include top-level as well as working-level scientists and technical personnel from all concerned constituencies.

2. NSSDC or ISO should establish interagency ties with other federal agencies and to the university community regarding dissemination of and access to pertinent data.

User Support

Information is not of much value unless it can be used. Much of the time the users of the stored data, although scientists or engineers, may not necessarily be skilled in the use of computers and communication networks. At archive centers providing data, staff consultants should be available to provide advice to potential and active users on methods of access, software, available utility programs, and methods of analysis. Satisfactory detailed documentation as well as general narratives must also be available.

Recommendations

1. NSSDC, in conjunction with ISO, should develop policy and guidelines for itself and the various archiving centers relating to the type of user support to be provided, and the means, necessary staff, method, and degree of support to be provided.

2. NSSDC, in conjunction with the various missions and programs, should establish a plan and schedule for a continuing series of Data Analysis Workshops to exchange information and to describe new and existing techniques.

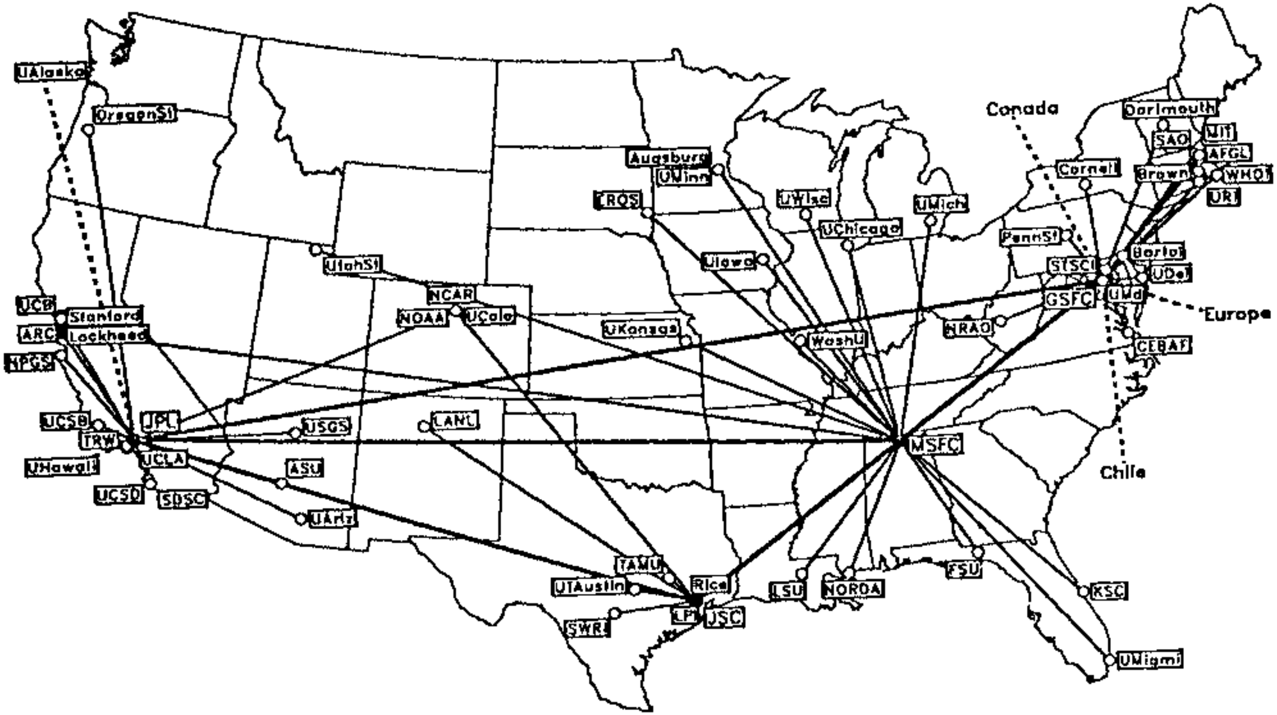
Appendix A

A SUMMARY OF SPAN

SPAN is the Space Physics Analysis Network. This important research tool of the NASA scientific community links space researchers from scores of institutions throughout the United States. The SPAN system is growing within the United States, and it also is expanding to connect NASA scientists with European and Japanese space research institutions.

The SPAN system serves many functions. Its paramount purpose is to provide scientists with a tool that improves their productivity. SPAN has traditionally been used to exchange mail messages, to send data back and forth for scientific papers and analysis workshops, and to share scientific software. SPAN has played a crucial role by disseminating spacecraft data in near-real time during several recent NASA and ESA programs, such as the International Cometary Explorer (ICE) spacecraft encounter with Comet Giacobini-Zinner (see "Behind the Scenes During a Comet Encounter" by J.L. Green and J.H. King, 2EOS1, March 4, 1986, p. 105), the Voyager 2 encounter with Uranus, and the Giotto spacecraft encounter with Comet Halley (see "Networking Ground-Based Images of Comet Halley during the Giotto Encounter" by D. Rees et al., 2EOS1, December 16, 1986, p. 1385). Of course, SPAN has served a variety of broader purposes. It has

Space Physics Analysis Network
North American Sites



demonstrated the value of intercommunication to a broad range of the user community. It has provided (and will continue to provide) an excellent testing ground for trying new technologies and for evaluating ideas about processing, storing, and transferring various kinds of information. Until the recent availability of computers sponsored by the National Science Foundation, the SPAN network provided one of the few opportunities for NASA researchers to have ready access to the supercomputers needed for large-scale numerical simulation of magnetospheric and ionospheric plasma systems. SPAN has also increased its usefulness substantially by the addition of the National Space Science Data Center (NSSDC) as a data node on the system. The network, which was originally based on a star-shaped configuration, has been redesigned over the last year to take advantage of NASA's new Program Support Communications Network (PSCN).

The redesign of SPAN takes advantage of the enhanced capabilities provided by PSCN and allows for the orderly and more efficient growth of the network. The new SPAN configuration uses four central routing nodes, located at NASA Goddard Space Flight Center (GSFC), NASA Marshall Space Flight Center (MSFC), NASA Johnson Space Center (JSC), and the Jet Propulsion Laboratory (JPL). These "routing centers" are linked by 56-kbps communication lines to form the SPAN "backbone." All other SPAN nodes will then be linked to the closest of these four routing centers by using 9.6-kbps lines (called tail circuits). The network will continue for the present to use the DECNET protocols that it has used in the past for its communications system; SPAN personnel are looking very closely at substituting a more general networking protocol as this becomes necessary and practical.

A new project management plan accompanies the new SPAN configuration. The Data Systems Users Working Group (DSUWG) as a whole determines the general technical direction in which SPAN will evolve and ensures that changes made in the network will improve the use of the system for scientific purposes. The DSUWG steering committee (which consists of the DSUWG chair and the subgroup chairs) approves all special mission-dependent uses of SPAN. Various project scientists coordinate the science activities and general network use, coordinate all SPAN activities with OSSA at NASA headquarters, coordinate with other NASA and ESA projects for SPAN use, and work with the project manager to develop funding projections to run the system. The project manager coordinates all SPAN activities with the DSUWG advisory committee, the project scientists, and the network managers; implements the advisory committee recommendations; and manages the evolution of SPAN while maintaining contact with other NASA projects that will have SPAN connections or gateways.

The network manager is responsible for day-to-day operational management of the network, working closely with the routing center managers and node managers to provide user support. The routing center managers handle all of the network lines coming into the respective routing centers, maintain and operate the SPAN router services, coordinate new network nodes coming into the routing center, and support mission-specific SPAN usage. The remote node managers maintain the network hardware and software at the node, maintain all connections with other local area networks, keep the network manager informed of node actions that will adversely affect the network, and ensure compatibility with SPAN.

The SPAN system has grown up as a grass roots effort by concerned and highly motivated scientists in the NASA space science community. It is continuing to grow dynamically and vigorously as the research potential of such a system becomes clear to more and more in the space science community, both in the United States and abroad. In all of its activities, SPAN is overseen by the people best suited to judge how it should be run, namely, the SPAN users themselves. This DSWUG continues to make SPAN responsive to the evolving needs of NASA's research community.

Appendix B

THE NASA SCIENCE INTERNET

The NASA Science Internet (NSI) program was initiated in 1986 in response to a widely perceived need for a NASA science network to support a broad range of science disciplines. Originally planned as a network based on the ARPA developed TCP/IP protocol suite and referred to as the NASA Science Network (NSN), it has been broadened in scope to include other protocols (primarily DECNET). The charter for the program is to provide nine networks capable of handling all of the OSSA Code E network requirements for both flight and pilot projects. Codes R and S have also expressed interest and support for this project.

The requirements for NSI have been developed by the PSCN Users Committee, with membership from NASA centers, Codes E and T and associated service groups, user representatives, and outside experts. This user group has been involved through all of the review and planning cycles, and user involvement is expected to continue at several levels through network deployment, use, and evolution.

The TCP/IP protocol suite was chosen because for its full range of network services across a broad mix of system architectures, its capability to provide interoperability for heterogeneous systems, and its mature set of implementations. The use of TCP/IP will permit easy interconnections with other agency networks (ARPA, NSF, DOE) that also use the same protocols.

The initial configuration of the network, which began in late spring 1987, is to have hub sites located at four major NASA centers (ARC, GSFC, JPL, and MSFC), all interconnected by one or more 56-kbps links. Tail circuits to major user sites will also be run at 56 kbps, although smaller sites or those with lower demands may use 9.6-kbps links.

The network topology will include placement of gateways at the hub locations and at each major user site. Each gateway is capable of handling multiple 56-kbps links or T1 links (1.544 Mbps) and of driving one or more Ethernet or other Local Area Network (LAN) interfaces. This model is the same as used by NSF in that connections are made to a site rather than to a single system.

More than 100 sites are in the queue for connection, spanning the disciplines of astronomy, climate, oceans, land, planetary, life sciences, and material science. The topology is flexible and for major concentrations of user activity at a distance from the hubs other subhubs may be established. Hub locations at JSC, KSC, and LeRC are also anticipated in support of a variety of research activities at those locations.

The network model assumes that each site or campus will provide the Ethernet (or other) connections between the site gateway and the user hosts. So many universities and other sites are installing such campus networks that this is both functional and cost effective. Any NASA sponsored researcher at a connected site may gain access to the gateway and hence to NSI itself. With the development of interagency agreements other users may also be permitted access.

The NSI project is working with Codes E and I to investigate cooperative agreements with other agencies (notably, NSF, ARPA, and NOAA) to share connectivity and links. By reaching such agreements, it is possible to gain greater connectivity among researchers and science resources in a cost-effective way. Many hundreds of major universities, labs, and other facilities are already connected to the ARPA Internet and the number of host systems is in the thousands. The connection of ARPANET to LAN's connecting multiple users also permits multiple, low-speed circuits to be aggregated and shared among users. This improves overall access and bandwidth and will often permit acquisition of higher speed links at a cost savings because of economies of scale in the tariff structure.

The NSI Project Office, which has been established at ARC, is responsible for the engineering, installation, operations, and maintenance of the network. This activity will be coordinated with the ISO Program Office, the PSCN Users Committee, and the PSCN Project Office at MSFC. An executive steering committee with appropriate user and management composition will provide the necessary direct guidance.

In addition to managing the network, the NSI Project Office expects to provide a number of user services. These will include user and site name services ("white pages") and facilities directory ("yellow pages") and other user aids. A number of different science resources such as supercomputers, data archives, research labs, and colleagues will be accessible directly over the Internet and they will be catalogued in the various directories. Site assistance and information services will also be provided in conjunction with the site coordinators to be identified at each user site.

The construction of an internet that supports TCP/IP and DECnet is to be an early activity of NSI. Individual networks based on these protocols already exist (ARPA, NSF, SPAN) but they do not interoperate except to transfer mail. Carrying both protocols on the same circuits, as is now often done on Ethernet is only part of the solution; it must be possible for machines using these different protocol suites to communicate with one another across the range of services. Within NSI there is some active work going on to solve this problem, and prototype services have already been constructed and are being experimented with.

The NSI program has recognized the need for long-range planning and incorporates a testbed activity, outside the normal operational systems, to support evaluation of new services and facilities. The protocol interoperability evaluation is one such activity. Others will include high-speed fiber LAN, higher-speed (T1 and up) backbones, circuit management to permit bulk traffic flows for image transfer, and planning for ISO/OSI migration. These new services will first be evaluated in testbeds before being placed in service in the operational networks.

The U.S. academic and network research communities are all using the same TCP/IP based protocols and this has significant advantages for NSI. Rather than have to develop a research and development activity within NASA, or to wait on a single vendor to supply the enhancements, there is an active national research program that is available as a resource. This is coordinated under the ARPA Internet Activities Board, in conjunction with ARPA and NSF funded research activities. Members of the PSCN Users Committee and other members of the community participate in these activities, thus ensuring that the needs of the NASA science community are well represented.

As NASA moves into the Space Station era networking will assume even greater value in support of the envisioned telescience activities. The NSI is expected to evolve to support these activities, thus taking advantage of early developments in the infrastructure. The importance of the NSI activity is being recognized in the Space Station era planning and in the planning for the Science Applications Information System (SAIS).