ED 340 594 SE 052 448

TITLE	Seeking Solution: High-Performance Computing for Science. Background Paper.
INSTITUTION	Congress of the U.S., Washington, D.C. Office of Technology Assessment.
REPORT NO	OTA-BP-TCT-77
PUB DATE	Apr 91
NOTE	50p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402-9325 ($2.25).
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Computer Centers; Computer Science; *Financial Support; *Government Role; *Research and Development; Technological Advancement
IDENTIFIERS	*High Performance Computing; *Supercomputers

ABSTRACT

This is the second publication from the Office of Technology Assessment's assessment on information technology and research, which was requested by the House Committee on Science and Technology and the Senate Committee on Commerce, Science, and Transportation. The first background paper, "High Performance Computing & Networking for Science," published in 1989, framed the outstanding issues; this background paper focuses on the federal role in supporting a national high-performance computing initiative. Chapter 1, "High-Performance Computing and Information Infrastructure for Science and Engineering," discusses the goals of the initiative, the government's role, the structure of federal policy, major strategic concerns, and long-range planning needs. Chapter 2, "Policy Considerations for High-Performance Computing," describes the difficulties and barriers to advancing computer technology, providing access to resources, and expanding and improving usage. The purposes of these centers are also discussed in this section. Chapter 3, "High-Performance Computers: Technology and Challenges," discusses the research and development process and the evolution of computer technology. Brief descriptions of national and other high-performance computer facilities are appended. (KR)

I-R

SEEKING SOLUTIONS: HIGH-PERFORMANCE COMPUTING FOR SCIENCE

BACKGROUND PAPER

CONGRESS OF THE UNITED STATES OFFICE OF TECHNOLOGY ASSESSMENT

SEEKING SOLUTIONS: HIGH-PERFORMANCE COMPUTING FOR SCIENCE

BACKGROUND PAPER

CONGRESS OF THE UNITED STATES OFFICE OF TECHNOLOGY ASSESSMENT

Recommended Citation:

U.S. Congress, Office of Technology Assessment, *Seeking Solutions: High-Performance Computing for Science—Background Paper*, OTA-BP-TCT-77 (Washington, DC: U.S. Government Printing Office, April 1991).

For sale by the Superintendent of Documents
U.S. Government Printing Office, Washington, DC 20402-9325
(order form can be found in the back of this background paper)

Foreword

High-performance ''supercomputers'' are fast becoming tools of international competition and they play an important role in such areas as scientific research, weather forecasting, and popular entertainment. They may prove to be the key to maintaining America's preeminence in science and engineering. The automotive, aerospace, electronic, and pharmaceutical industries are becoming more reliant on the use of high-performance computers in the analysis, engineering, design, and manufacture of high-technology products.

Many of the national and international problems we face, such as global environmental change, weather forecasting, development of new energy sources, development of advanced materials, understanding molecular structure, investigating the origin of the universe, and mapping the human genome involve complex computations that only high-performance computers can solve.

This is the second publication from our assessment on information technology and research, which was requested by the House Committee on Science and Technology and the Senate Committee on Commerce, Science, and Transportation. The first background paper, *High Performance Computing & Networking for Science*, published in 1989, framed the outstanding issues; this background paper focuses on the Federal role in supporting a national high-performance computing initiative.

OTA gratefully acknowledges the contributions of the many experts, within and outside the government, who served as panelists, workshop participants, contractors, reviewers, and advisors for this document. As with all OTA reports, however, the content is solely the responsibility of OTA and does not necessarily constitute the consensus or endorsement of the advisory panel, workshop participants, or the Technology Assessment Board.

John H Gibbons

JOHN H. GIBBONS
Director

High-Performance Computing and Networking for Science Advisory Panel

John P. (Pat) Crecine, *Chairman*
President, Georgia Institute of Technology

Charles Bender
Director
Ohio Supercomputer Center

Charles DeLisi
Chairman
Department of Biomathematical Science
Mount Sinai School of Medicine

Deborah L. Estrin
Assistant Professor
Computer Science Department
University of Southern California

Robert Ewald
Vice President, Software
Cray Research, Inc.

Kenneth Flamm
Senior Fellow
The Brookings Institution

Malcolm Getz
Associate Provost
Information Services & Technology
Vanderbilt University

Ira Goldstein
Vice President, Research
Open Software Foundation

Robert E. Kraut
Manager
Interpersonal Communications Group
Bell Communications Research

Lawrence Landweber
Chairman
Computer Science Department
University of Wisconsin-Madison

Carl Ledbetter
President/CEO
ETA Systems

Donald Marsh
Vice President, Technology
Contel Corp.

Michael J. McGill
Vice President
Technical Assessment & Development
OCLC, Computer Library Center, Inc.

Kenneth W. Neves
Manager
Research & Development Program
Boeing Computer Services

Bernard O'Lear
Manager of Systems
National Center for Atmospheric Research

William Poduska
Chairman of the Board
Stellar Computer, Inc.

Elaine Rich
Director
Artificial Intelligence Lab
Microelectronics & Computer Technology Corp.

Sharon J. Rogers
University Librarian
Gelman Library
The George Washington University

William Schrader
President
NYSERNET

Kenneth Toy
Postgraduate Research Geophysicist
Scripps Institution of Oceanography

Keith Uncapher
Vice President
Corporation for the National Research Initiatives

Al Weis
Vice President
Engineering & Scientific Computer
Data Systems Division
IBM Corp.

NOTE: OTA appreciates and is grateful for the valuable assistance and thoughtful critiques provided by the advisory panel members. The panel does not, however, necessarily approve, disapprove, or endorse this background paper. OTA assumes full responsibility for the background paper and the accuracy of its contents.

OTA Project Staff—High-Performance Computing

John Andelin, *Assistant Director, OTA*
Science, Information, and Natural Resources Division

James W. Curlin, *Telecommunication and Computing Technologies Program Manager*

Fred W. Weingarten,[1] *Project Director*

Elizabeth I. Miller, *Research Assistant*

Administrative Staff

Elizabeth Emanuel, *Office Administrator*

Karolyn St. Clair, *Secretary*

Jo Anne Price, *Secretary*

[1] Through June 30, 1990.

Contents

Boxes

Figure

Table

Chapter 1

High-Performance Computing and Information Infrastructure for Science and Engineering

Introduction

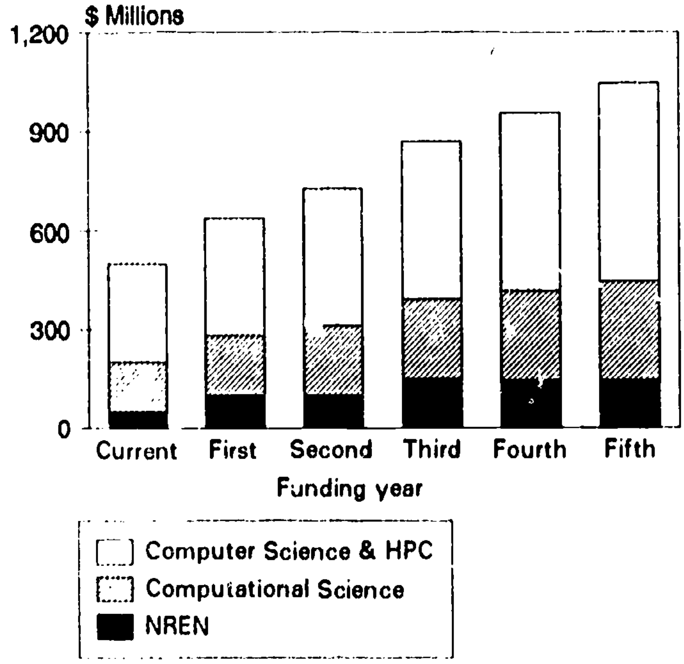
Information technology is a critical element for science and engineering. The United States is building a nationwide computer-communication infrastructure to provide high-speed data services to the R&D community, but the mere installation of hardware is not enough. Whether very fast data communication networks and high-performance computers deliver their promised benefits will depend on the institutions, processes, and policies that are established to design and manage the new system.

OTA was asked by the House Committee on Science, Space, and Technology and the Senate Committee on Commerce and Transportation to examine the role that high-performance computing, networking, and information technologies are playing in science and engineering, and to analyze the need for Federal action. An OTA background paper, released in September 1989, explored and described some key issues.[1] This background paper examines high-performance computing as part of the infrastructure proposed in the Office of Science and Technology Policy (OSTP) initiative. A detailed OTA report on the National Research and Education Network (NREN) is scheduled for release later in 1991.

Six years ago, Congress directed the National Science Foundation (NSF) to establish an Advanced Scientific Computing Program designed to increase access by researchers to high-performance computing. That program resulted in the establishment of five national centers for scientific supercomputing. Since then, one of the centers has been left unfunded; but the other four are still operating.

During the last 5 years, legislation has been introduced in Congress calling for the establishment of a high-capacity, broadband, advanced national data communications network for research. Over the years, congressional interest has grown as this concept has evolved into a plan for an integrated national research and education network (NREN) consisting of an advanced communication network linked to a variety of computational and information facilities and services.

Figure 1—Estimated Proposed Funding Levels for Federal High-Performance Computing Program



SOURCE: Office of Science and Technology Policy, *The Federal High-Performance Computing Program* (Washington, DC: Office of Science and Technology Policy, 1989), app. C, p. 26.

In September 1989, at the request of Congress, OSTP submitted a draft plan developed by the Federal Coordinating Council for Science, Engineering, and Technology (FCCSET). The plan called for a "National High-Performance Computing Initiative" that includes both a national network and initiatives to advance high-performance computing (see figure 1). In testimony to the 101st Congress, the director of OSTP stated that this Initiative was among the top priorities on the science agenda. On June 8, 1990, the National Science Foundation announced a $15.8 million, 3-year research effort aimed at funding 5 gigabit-speed testbed experimental networks. These test networks are the first step in developing a high-speed nation-

[1] U.S. Congress, Office of Technology Assessment, *High Performance Computing & Networking for Science*, OTA-BP-CIT-59 (Washington, DC: Government Printing Office, September 1989).

wide broadband advanced communication network in collaboration with the Defense Advanced Research Project Agency (DARPA).

OSTP set forth its plans for a Federal High-Performance Computing and Communications Program (HPCC) in a document released on February 4, 1991, supporting the President's Fiscal Year 1992 budget.[2] The Program proposes to invest $638 million in fiscal year 1992, an increase of about 30 percent over the 1991 level. These funds will support activities in four program areas: 1) high-performance computing systems; 2) advanced software technology and algorithms; 3) National Research and Education Network; and 4) basic research and human resources.

High-Performance Computing: A Federal Concern

Concern about information technology by high-level policymakers is a recent phenomenon. Researchers who see the importance of data exchange have managed to secure funding for computers and communications out of the limited Federal agency research budgets. Agencies such as DARPA and NSF have quietly developed computer and network-related programs without major administration initiatives or congressional actions. But the atmosphere is now different for several reasons.

First, researchers cannot consistently obtain needed information resources because of the cost. High-end scientific computers cost several million dollars to purchase and millions more per year to operate. Universities grew reluctant in the late 1970s and early 1980s to purchase these systems with their own funds, and government investment in computers slowed. In the meantime, researchers learned more about the use of high-performance computing. The machines became more powerful, doubling in speed about every 2 years. The scientific community slowly became aware of the lost opportunities caused by lack of access to high-performance computers and other powerful information technologies.

Second, information resources—computers, databases, and software—are being shared among disciplines, institutions, and facilities. These are being linked as common resources through networks to users at desktop workstations. A need has grown for better coordination in the design and operation of these systems; this will be particularly important for a national data communications network.

Third, although the U.S. computer industry is relatively strong, there is concern about increasing competition from foreign firms, particularly Japanese. Over the last decade, the Japanese Government has supported programs, such as the Fifth Generation Project (it is now planning a Sixth Generation initiative) and National Superspeed Computer Project, designed to strengthen the Japanese position in high-performance computing. During the last 2 years there have been difficult trade negotiations between the United States and Japan over supercomputer markets in the respective countries. This has raised concern about the economic and strategic importance of a healthy U.S. high-performance computing industry.[3]

Fourth, concern for the Japanese challenge in high-performance computing goes beyond the competitiveness of the U.S. supercomputer industry. Computational simulation in engineering design and manufacturing is becoming a major factor in maintaining a competitive posture in high-technology industries such as automotives, aerospace, petroleum, electronics, and pharmaceuticals. It is in the availability and application of high-performance computing to increase productivity and improve product quality where the greatest future economic benefits may lie.

Finally, the infrastructure of this interlinked set of technologies is considered by some to be a strong basis for the development of a universal broadband information system. A very high-capacity digital communication network and information services, as visualized, would carry entertainment, educational, and social services to the home and support a broad range of business and education services. A

[2] OSTP, *Grand Challenges: High-Performance Computing and Communications*, (Washington, DC: OSTP, 1991), p. 57.

[3] High-performance computers are considered to be strategically important to the United States because of the central role computers play in the economy, security, manufacturing, and research and development. *See* Office of Science and Technology Policy, *The Federal High Performance Computing Program* (Washington, DC: Office of Science and Technology Policy, 1989), pp. 12-13. *See* U.S. Congress, Office of Technology Assessment, *Making Things Better: Competing in Manufacturing*, OTA-ITE-443 (Washington, DC: U.S. Government Printing Office, February 1990), p. 241, for a comprehensive view of the U.S. competitive position in manufacturing, including computer-related industries. An assessment of the status the U.S. supercomputing industry will be included in a forthcoming OTA report on Japan and International Trade.

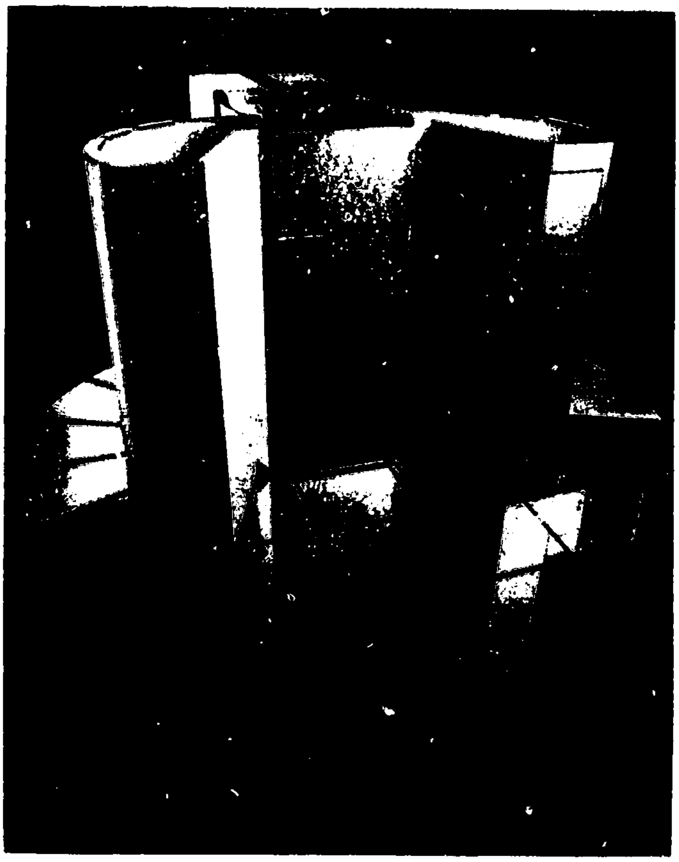


Photo credit: Cray Research, Inc.

The CRAY Y-MP/832 computer system is the top-of-the-line supercomputer of Cray Research, Inc. It contains 8 central processors and 32 million 64-bit words of memory.

nationwide network for research and education could be a starting point, and could gradually broaden to this vision.

Multiple Goals for an Initiative

The supporting arguments for a Federal High-Performance computing/networking initiative center on three objectives:

1. *To advance U.S. research and development* critical to U.S. industry, security, and education by providing researchers with the most powerful computers and communication systems available. This objective is based on a vision of computers and data communication technologies forming a basic infrastructure for supporting research. This goal has been proposed in several reports and policy papers.[4]
2. *To strengthen the U.S. computer industry* (particularly the high-performance computers and high-speed telecommunications) by testing new system concepts and developing new techniques for applications. Federal Coordinating Council for Science, Engineering, and Technology (FCCSET) and the Institute of Electrical and Electronics Engineers, among others, strongly endorse this view.[5]
3. *To enhance U.S. economic and social strength* by stimulating the development of a universal information infrastructure through development of new technologies that could serve as a system prototype.[6]

Strong sentiment exists among some Members of Congress for each of these three objectives. Furthermore, the goals are closely related and nearly inseparable—most discussion and proposals for computing and networking programs reflect elements of all three.

Not everyone in Congress or the executive branch agrees that all goals are equally important or even appropriate for the Federal Government. Some consider the current level of government spending to advance scientific knowledge to be adequate, and they believe that other needs have higher priority. Others point out that since information technology is now central to all R&D, it is important to create a modern information infrastructure in order to realize the benefits from government investment in science and engineering.

[4] Peter D. Lax, "Report of the Panel on Large-Scale Computing in Science and Engineering" (Washington, DC: National Science Foundation, 1982), p. 10. EDUCOM, Networking and Telecommunications Task Force, *The National Research and Education Network: A Policy Paper* (Washington, DC: EDUCOM, 1990) p. 3.

"The goal of the National Research and Education Network is to enhance national competitiveness and productivity through a high speed . . . network infrastructure which supports a broad set of applications and network services for the research and instructional community."

[5] Executive Office of the President, Office of Science and Technology Policy, *The Federal High Performance Computing Program* (Washington, DC: September 1989), p. 1.

"[A goal of the High Performance Computing Program is to] maintain and extend U.S. leadership in high performance computing, and encourage U.S. sources of production."

[6] *Congressional Record* comments on introduction of bill. For example, Senator Gore stated the following, when introducing his bill, S1067:

"The nation which most completely assimilates high performance computing into its economy will very likely emerge as the dominant intellectual, economic, and technological force in the next century.

U.S. industry must produce advanced yet economical systems which will meet the needs of users found in each of the major sectors. . . . If this is not done by U.S. Government leadership, it will be done by foreign leadership to the detriment of U.S. national interests."

Some disagree with an initiative that resembles "industrial policy"—i e , policy aimed at supporting specific private sector enterprises. They argue that the government should not intervene to support either the supercomputer or the telecommunications industry. Proponents of government intervention argue that the dominant position of the U.S. supercomputer industry has historically resulted from heavy Federal investments in computing for research and that the future health of the industry will require continued Federal attention.

Some ask why science should get early preferred access to what ultimately may become a universal communication service, and suggest that selectively providing such resources to science might delay broader adoption by the public that promises even greater payoffs. They are also wary of the government providing or subsidizing telecommunication services that should, in their view, be provided by the private sector. They argue that a universal network is best achieved through the expertise and resources of the commercial communication and information industries. Proponents of Federal action maintain that the science network will be an important prototype to develop and test new standards and technologies for extremely high-speed packet-switched data communication. Furthermore, in their view, a network oriented to research and education would be a valuable testbed for developing applications and better understanding how a universal network would be used.

These debates reflect in part different philosophies, values, and expectations about future events that must be resolved in a political process. The assumptions underlying each of these three goals—advance U.S. R&D, strengthen the U.S. computer and telecommunications industry, and enhance U.S. economic and social strength—are generally soundly based because:

1. **Scientific users need access to advanced computers, communication systems, databases, and software services**—Scientific and engineering research in the United States cannot retain its world-class position without the best available information and communication technologies. These include advanced computer systems, very large databases, and high-speed data communications, local workstations, electronic mail service, and bulletin boards. Such technology does not simply enhance or marginally improve the productivity of the research process; it enables research that could not be performed otherwise. Simulating the complex behavior of the Earth's climate, analyzing streams of data from an Earth satellite or visualizing the interactions of complex organic molecules are impossible without these new technologies. Furthermore, many more important applications await the as-yet-unrealized capabilities of future generations of information technology.
2. **Major Federal research applications have stimulated the computer industry and will likely continue to do so**—Scientific and engineering applications have stretched the capacities of information technologies and tested them in ways that other applications cannot. Eventually, the techniques and capabilities developed to serve these demands make their way into the broader community of computer users. Although the computer industry structure and markets are changing, this form of technology transfer will likely continue.
3. **U.S. economic growth and societal strength can be assisted by the development of a national information infrastructure that couples a universal high-speed data communication network with a wide range of powerful computational and information resources**—In a recent report, *Critical Connections: Communication for the Future*, OTA stated:

> Given the increased dependence of American businesses on information and its exchange, the competitive status among businesses and in the global economy will increasingly depend on the technical capabilities, quality, and cost of [their] communication facilities . . . Failure to exploit these opportunities is almost certain to leave many businesses and nations behind.[7]

In that report, OTA listed "modernization and technological development of the communication infrastructure" as one of the five key areas of future policy concern. The high-performance computing

[7] U.S. Congress, Office of Technology Assessment, *Critical Connections: Communication for the Future*, OTA-CIT-407 (Washington, DC: U.S. Government Printing Office, January 1990), p. 6.

and networking initiative reflects a mixture of three basic goals: 1) enhancing R&D, 2) accelerating innovation in U.S. information technology, and 3) stimulating the development of a universal broadband digital network in the United States. Achieving the last goal will ultimately bring information and educational opportunities to the doorstep of most American homes.

An Infrastructure for Science

Science and Information Technology Are Closely Linked

Whether computers are made of silicon chips, optical glass fibers, gallium arsenide compounds, or superconducting ceramics, and regardless of the architecture, the basic elements of computers are the same—data, logic, and language.

- *Data* is the substance that is processed or manipulated by the technology; often, but not always, numerical.
- *Logic* is the nature of the process, from basic arithmetic to extremely complex reasoning and analysis.
- *Language* is the means of communicating from the user to the machine what is to be done, and from the machine to the user the result of that action.

These three elements are also basic to science. They characterize the nature of research and the work of scientists.

- Researchers collect data from measurements of natural phenomena, experiments, pure mathematics, and, increasingly, from computer calculations and simulations. Data can take many forms, e.g., numbers, symbols, images, sounds, and words.
- Researchers build logical structures—theories, mathematical and computer models, and so on—to describe and understand the phenomena they are studying.
- Researchers communicate their work among themselves in common scientific languages. This communication is a continuing process—both formal and informal—that lies at the heart of the scientific method. It is based on exposing ideas to the critical review of peers, allowing the reproduction of experiments and analyses, and encouraging the evolution of understanding based on prior knowledge.

Scientists invented the computer to serve research needs during World War II. In the late 1960s, research needs led to the development of ARPANET —the first nationwide communication system designed specifically to carry data between computers. NSF operated a Computer Facilities Program in the late 1960s and early 1970s that assisted universities in upgrading their scientific computing capabilities for research and education. Today, computers are used throughout society, but researchers, joined by industry, are still driving the evolution of information technology and finding new applications for the most powerful computer and communication technologies.

The invention of the printing press in the 15th century created the conditions for the development and flourishing of modern science and scholarship. Not only did the press allow authors to communicate their ideas accurately, but the qualities of the medium stimulated entirely new methods and institutions of learning. Similarly, electronic information technology is again changing the nature of basic research. The character of the research—the way data are collected, analyzed, studied, and communicated—has changed because of technology.

Computational research has joined experimentation and theory as a major mode of investigation. Scientists now use computer models to analyze very complex processes such as the flow of gases around a black hole or the wind patterns around the eye of a hurricane or typhoon (see boxes A and B). These and other areas of research, such as global climate change, can be accomplished only with high-performance computing. They cannot use conventional mathematical and experimental approaches because of the complexity of the phenomena.

Research is generating data at unprecedented rates. The human genome database is projected to eventually contain over 3 billion units of information. Earth observation experiments in space will collect and send to Earth trillions of units of data daily. A single image of the United States, with resolution to a square yard, contains nearly a trillion data points. Current data storage technologies are unable to store, organize, and transmit the amounts of data that will be generated from these projects. Long-term storage capabilities must be researched and developed. "Big science" projects, such as those mentioned above, should devote a portion of their budgets for R&D in high-capacity data storage systems if such large projects are to be successful. New forms of institutions and procedures to manage massive data banks are also needed.

Journal articles have been the major form of communication among scientists. Publishers are beginning to develop electronic journals, accessed directly over communication networks or distributed in computer readable form. The different nature of electronic storage means that these new "publications" will likely look, behave, and be used differently than printed journals. They may contain information in a variety of forms: high-definition video, moving images, sound, large experimental data sets, or software. Using so-called "hypermedia" and multimedia techniques, these electronic journals can be linked to other related articles, films, and so on. They can evolve and change over time, containing later annotations by the original author or others, or references to later articles that advanced or stemmed from the original work.

Scientists communicate with each other continually by letter, telephone, conferences, and seminars, and by meeting personally around the departmental coffee pot. These modes of communication are often as important to research as formal publications. All of these modes will likely continue in some form, but digital communication systems provide many new powerful ways to communicate—electronic mail, computer conferences, and bulletin boards. Proximity, time, and travel are less important in using these new communication paths. With bulletin boards and electronic mail, information can be exchanged much faster than by mail and with accuracy and detail that is impossible to achieve with telephone. Since the participants need not travel, computer conferences can accommodate large numbers of participants. Sessions can take place over weeks or months, and people can participate in the electronic meeting wherever they are at whatever time is convenient.

Box A—Black Holes: The Mysteries of the Universe

A black hole is an object in space, whose mass is so dense that nothing is able to escape its gravitational pull, not even light. Astronomers think the universe is populated with black holes that are the remains of collapsed stars. Much of the research conducted on the universe has implications for other areas of study, such as physics. Below is a visualization of the three dimensional flow of gases past a black hole. The computer codes used to create this image can also be used to determine the accuracy of computer generated models of three dimension fluid flows.

Collaborative efforts between two NSF-funded supercomputer centers, the National Center for Supercomputing Application (NCSA) in Illinois and The Cornell Theory Center, resulted in a video of the phenomenon pictured. The computer code used to derive the data was written by researchers at Cornell and was run on Cornell's IBM 3090 computer. NCSA remotely accessed the data via NSFNET from Cornell. At the Illinois center researchers worked with a scientific animator who, using Waverfront Technologies Graphic Software tools and a graphics packaged designed at NCSA, processed the data on a Alliant computer. The research team created a rough contoured image of the cell with the Alliant. The researchers returned to Cornell and modified the contouring graphics programs, creating a videotape on Silicon Graphics workstation at Cornell. The project was the first joint effort between NSF supercomputer centers. Utilizing the expertise of two centers was instrumental in graphically depicting three-dimensional fluid flow.

A National Infrastructure for Research and Education

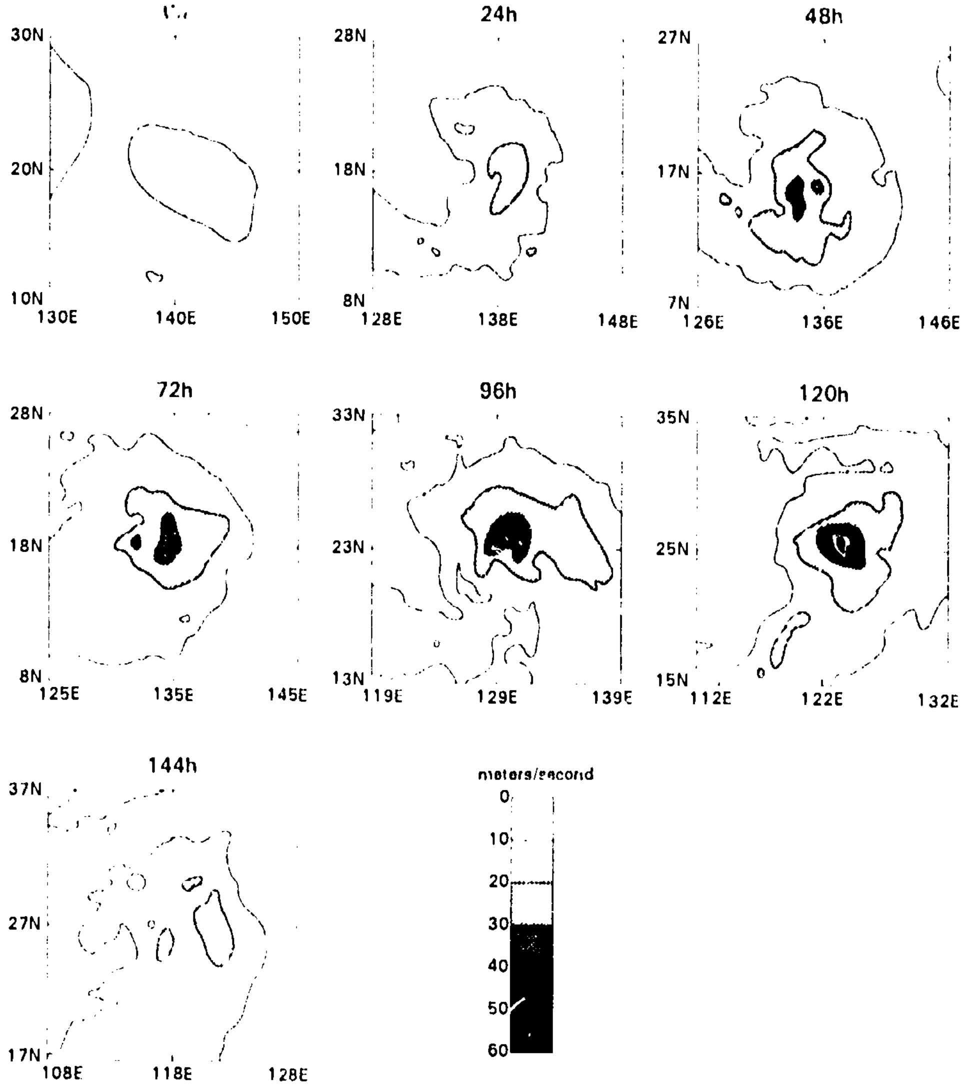
These information technologies and applications are merging into an interconnected network of resources referred to in telecommunication's parlance as an "Information Infrastructure." This infrastructure is a conceptual collection of linked resources made up of:

Box B—Supertyphoon Hope

A typhoon is a tropical storm confined to the western Pacific Ocean, referred to as hurricanes in the Western Hemisphere. The cyclonic storms are usually accompanied by extremely low atmospheric pressure, high winds of over 100 knots, and vast amounts of rainfall. These storms can wreak havoc when they reach land, at which time they dissipate. A series of computer-generated images that trace the 6-day evolution of supertyphoon Hope are shown here. The supertyphoon's course was simulated using a computer model. Researchers were able to measure the precision of their model by comparing its results to the data gathered during the storm in 1979. Developing accurate weather models continues to be difficult despite advances in technology. Weather models must incorporate many variables such a winds, temperatures, the oceans, and atmospheric pressures.

Researchers processed their weather model at the National Center for Atmospheric Research (NCAR) remotely from their home institute, Florida State University (FSU). Using a NASA computer network, the group accessed an IBM 4381 computer, which served as the front end machir for a Cray X-MP. After the data were processed at NCAR, it was transferred to magnetic tapes and mailed to FSU for further analysis. (An increase in bandwidth now allows the researchers to separate large data sets into sections and send them over high bandwidth computer networks.) At FSU the data were translated into images using a Silicon Graphics Workstation. Data collected from the storm in 1979 are stored on computers at NCAR. The data were used to measure the accuracy of FSU's weather model. The computer-generated storm was accurate within hours of the actual events of Supertypoon Hope.

The facilities at NCAR were especially suited for the needs of the FSU researchers since both the staff and resources were geared towards atmospheric and ocean modeling. Researchers frequently visit NCAR for user conferences and have become familiar with many of the technical support staff.



A Nationwide High-Speed Broadband Advanced Information Communication Network

This computer-to-computer network is composed of many parts—local networks on campuses and in research facilities, State and regional networks, and one or more national "backbone" networks interconnecting them all. This domestic backbone would link to networks in other countries. Some of the domestic networks will be private commercial networks; others may be operated by private nonprofit organizations, and still others may be government-funded and/or managed.

Specialized and General Purpose Computers

Users will be able to access the newest, most powerful supercomputers. There will be a variety of specialized machines tailored to specific uses and applications because of the developing nature of current computer architectures. They will be used for database searches, graphical output, and artificial intelligence applications such as pattern recognition and expert systems. Researchers will have access to a heterogeneous computing environment where several specialized machines, each with their own strengths, are linked through a software network that will allow users to simultaneously exploit the power of each computer to solve portions of a single application.

Collections of Specialized Applications Programs

Some application programs are extremely large and represent years of development effort. They may be maintained and updated centrally and made available over the network. Groups can also make available libraries of commercial or public domain software that could be distributed to local computers.

Remote Access to Research Instruments

Some research facilities house one-of-a-kind instruments, unique because of their cost or their site location, such as telescopes, environmental monitoring devices, oceanographic probes, seismographs, space satellite-based instruments, and so on. Remote use and control of these instruments is possible through the network infrastructure.

Services To Support and Enhance Scientific Communication

These services include electronic mail and conferencing systems, bulletin boards, and electronic journals through which researchers can communicate with each other. They also will include scholarly bibliographic reference and abstracting services, and online card catalogs linked to key research libraries.

"Digital Libraries" and Archives

These resources would contain collections of reference materials—books, journals, sound recordings, photographs and films, software, and other types of information—archived in digital electronic form. These also include major scientific and technical databases such as the human genome database, time-series environmental data from satellites, and astronomical images. Some visionaries see the network eventually providing access to a connection with a "Global Digital Library," a distributed collection of facilities that would store electronically most of the world's most important information.

Facilities for Analyzing and Displaying the Results of Computations

Researchers who simulate large, complex systems must develop ways to interpret these simulations to replace examining enormous quantities of computer-generated numbers. These researchers need to interact with the model; directly controlling the computer is the next step. Researchers are developing new ways to "see" their models by visualizing them directly or by use of methods such as holograms that provide three-dimensional views. Other researchers are developing tactile systems that, through special gloves and visors, allow a person to "feel" simulated objects or act as if they were moving about in a simulated environment ("virtual reality"). Researchers on the network will draw on specialized centers of technology and expertise to help them develop interfaces with computations and databases.

The Current Federal Picture

Recent Studies

Over the last 8 years, the key Federal science agencies and private groups have assessed the role of information technology in science and engineering research. These studies have concluded that computers and communication technology are critically important to R&D and the competitive position of the United States in the global economy. The studies pointed out shortcomings in the current system and recommended Federal actions.

The Lax Report

In 1982, the Panel on Large Scale Computing in Science and Engineering issued a report that became known as the "Lax Report" after its chairman, Peter Lax. It was jointly funded by the NSF and the Department of Defense. The panel noted that the U.S. research establishment seriously lacks access to high-performance computing. It found that this deficiency harms U.S. preeminence in R&D and threatens the current strong position of the U.S. computer industry. To remedy this, the panel recommended a national supercomputer program consisting of four basic components:

1. establish national supercomputing centers and develop "a nation-wide interdisciplinary network through which users will have access to facilities";
2. support research in software and algorithms—particularly work on parallelism and other new computer architectures for high-performance computing in the future;
3. support education and training programs for new users in order to assist the research community in using supercomputer applications; and
4. support research aimed at developing new, faster, supercomputers.

Variation of these four elements are repeated in the subsequent proposals and initiatives.

The Bardon/Curtis Report

At the request of Congress, NSF undertook an internal review of the Lax Report designed to form a program plan. The 1983 report, which became known as the Bardon/Curtis Report, offered an ambitious program plan with several recommendations for NSF:

1. "greatly increase its [NSF's] support for local computing facilities, including individual workstations, systems for research groups, specialized computer facilities, and local area networks";
2. establish 10 supercomputer centers;
3. support the establishment of networks to link users with the supercomputer centers; and
4. "support a program of academic research in the areas of advanced computer systems design."

National Academy of Sciences/COSEPUP

In 1989, the Committee on Science, Engineering, and Public Policy of the National Academy of Sciences (NAS) published a report, *Information Technology and the Conduct of Research*,[8] which examined the needs of science for new technological initiatives. This report, prepared by a panel chaired by Donald Langenberg, emphasized the changing form of research and its increased dependence on new information technologies. The report advised against leaving the design and operation of these programs only to the technical experts. Systems designers must learn what the users need, then design the system. The panel made two recommendations to do this:

- "The institutions that support U.S. research, including universities, industry, and Government should develop and support policies, services, and standards that help researchers use information technology more widely and productively, and
- "The institutions supporting the nation's researchers, led by the Federal government, should develop an interconnected national information technology network for use by all qualified researchers."

The panel recognized that industry, the universities, libraries, professional societies, and the Federal Government share responsibilities for this. The Federal role, according to the committee, should include leadership and coordination in the development of technologies and services to support the research and education needs in addition to funding.

National Association of State Universities and Land Grant Colleges (NASULGC)

In 1989, NASULGC issued a report, *Supercomputing for the 1990's: A Shared Responsibility*, on the need to make high-performance computing available for academic research, that contained recommendations for the Federal Government and universities. It points out that a computing infrastructure would have to include facilities operated by a variety of institutions beyond the Federal Government. Federal policy, it suggests, should be tailored to encouraging and leveraging private, regional, and local efforts. Recommendations for Federal action include:

[8] NAS report.

- supporting the national supercomputer centers and maintaining them at the technological leading edge;
- "fostering and encouraging university, state, and icgional supercomputing facilities" (of which the report identified 27);
- supporting the development of a national network; and
- assuring the "constancy" of support.

These recommendations reflect concern that the support of NSF national centers would draw funds away from computing at non-NSF centers. The report notes that the non-NSF centers will also play an important role in the future of scientific computing. NASULGC further observes that changing policies and unpredictable funding disrupts operations and discourages the development of facilities.

EDUCOM

EDUCOM is an association of higher education institutions. It functions as a clearing house for information and expertise about computers and communication technologies. EDUCOM's university consortium created and manages BITNET, a private, shared network that serves the networking needs of academics at a low cost. Its Networking and Telecommunications Task Force examined the Federal networking and computing initiatives and has produced several policy statements and reports.

A statement, released in March 1990, focuses on the network. It makes a series of specific recommendations on implementing the NREN, but its statement of the basic goal for a network is broad:

> [NREN]... is to enhance national competitiveness and productivity through a high-speed, high-quality network infrastructure which supports a broad set of applications and network services for the research and education community.[9]

Common Themes

The series of reports strikes some common themes. Three points, in particular, are important to the current policy debate.

First, the network has become the key element. Seen first as simply a means to access expensive oı highly specialized computing resources (similar to the initial intentions for ARPANET), the network has become the basic foundation for the information infrastructure, connecting researchers and students not only to computers, but providing access to a wide range of services.

The network is actually an internet, a family of networks (networks within networks), the design and operation of which needs coordination and leadership from the Federal science agencies. OTA's forthcoming report on the NREN will explore these issues in depth.

Second, educational needs are now part of the NREN plan although it is undecided how wide the range of users and institutions will be. In any event, this will affect network architecture and operating policies. Once referred to as a National Research Network (NRN), it is now known as a National Research and Education Network (NREN). This evolution was natural. It is impossible to separate education from research at the graduate level. There are also strong arguments for including undergraduates, secondary schools, and even primary schools in the system. To better coordinate the educational community's views on how the NREN may assist education, the Department of Education and State and local educators must be actively involved in the policy process.

The question of scope of the network extends to research in non-science scholarly disciplines, some of which are not well-funded by Federal programs. The wide range of services offered, including access to libraries, bibliographic services, electronic mail, bulletin boards, computer conferencing, and so on, extends the network's potential scholarly beneficiaries beyond just scientists and engineers.

A third comn.ıonly raised issue in the reports is the need to look beyond mere hardware. Computers need software that make them accessible and usable by researchers. A network needs software tools and data-bases that allow scientists to communicate effectively with one another. Databases need inquiry systems (search engines), indices, tables of contents, directories, hypertext, and other tools to enable users to search, identify, and retrieve the information they need for their work. To properly develop an infrastructure that is useful to all science, attention must be paid to software as well as hardware.

Other groups also rely on access to scientific information. Public interest groups with concerns of

[9] EDUCOM, The National Research and Education Network, op. cit., footnote 4, p. 3.

public safety and health, the environment, energy, defense policy, and so on, rely on access to scientific publications and databases and attendance at conferences and seminars. The press, particularly the specialized scientific and technical press, must access conferences, journals, and other forms of electronic communication.

In most cases, these applications, which enable effective access and use of information, are neither simple nor obvious. Developing them will require significant research and software development as well as better understanding of how information technology can best assist scholars in their work. The answers to these questions will also depend on the nature and breadth of the constituency for the network. Different users will have different skills, analytical strategies, and research styles depending, in part, on the traditions of their particular disciplines and their level of training.

The Government's Role Is Changing

The Federal Government has major responsibilities for the health of basic and academic research in the United States, both as a user of the products and from its role in supporting science and engineering to advance the economy and improve the qual:'y of life. The government is already participating heavily in the development and management of the existing R&D information infrastructure—in using it, funding it, and in setting policies. The government must deal with additional responsibilities resulting from the new infrastructure. The challenge will be to organize and assemble a government entity to: 1) identify and determine promising technological directions for the high-performance computing infrastructure; 2) evaluate progress over the course of the High-Performance Computing (HPC) initiative; and 3) make course corrections at the appropriate times.

First, facilities need to be highly interconnected. They must connect physically and logically with the network. Digital signals must conform to standards and protocols in the same way that electrical appliances must plug into standard 110 volt, 60 Hz AC power outlets. Users must be able to transmit and receive communications, programs, and data seamlessly and transparently to and from each nook and cranny in the system. Government policies and programs must be coordinated if interconnectability is to be achieved.

Second, many of the shared resources cut across agencies, institutions, disciplines, and programs. This feature of sharing is most obvious in the physical network; but it is also true for many of the computing facilities, data archives, and network services such as directories, bulletin boards, and electronic mail. Thus, many policy decisions regarding the use and access to these resources and services must be made at an interagency level. Furthermore, new private networks and service corporations now provide networking services to public and private customers. The policies of these private entities will become more important as the network moves towards full commercialization.

Third, the facilities will be expensive and require large capital investments. The Federal Government will be asked to share these costs with States, local governments, other countries, research and educational institutions, industrial users, service providers, and individual users. Private entities may be expected to contribute substantially as well. But while technological risks may be acceptable to private companies. The commercial risks may be unacceptable without government support.

Fourth, many of the resources on the network will be unique and of great national—even international—importance, e.g., a supercomputer dedicated to global climate modeling or the human genome database. Access to these scarce resources must reflect a cooperative set of goals, to determine access, decide who can use it, and to set national priorities. Federal policies are needed to balance and ensure equitable access and security. The Executive Office of the President (OSTP) and congressional committees may be called on to referee conflicts among competing interests from time to time. A well-organized Federal management system responsible for policy oversight and operation of the HPC and network infrastructure can anticipate or avoid many problems, and thus reduce the need for political resolutions.

Finally, the government must assist in advancing the state of computer and communication technologies to hasten the development of more powerful high-performance computers, faster data communications, and more effective software. Several studies by NAS, OSTP, and others have identified "Grand Challenges in Research" of critical national importance, but which are currently unachievable because of inadequate computing power.[10] What is needed are computers that are hundreds—even thousands—of times faster than the best now available. Similarly, the "data explosion" demands better and faster storage technology to archive large data sets. The rapidly growing communication needs of science require switched wide-area digital communication networks capable of moving billions of units of information per second to and from researchers.

The Structure of Federal Policy

Researchers foresee computers that will soon perform a trillion arithmetic steps ("teraflops") per second, data communication systems that can transmit billions of units of data ("gigabits") per second, and electronic storage systems that can store correspondingly large amounts of information and absorb and disgorge it at rates matched to the speed of the computers and communication lines. New hardware will require new streamlined software to operate the high-speed computers and communications networks efficiently.

Developing, Managing, and Funding Major Resources

A striking trend in information technology is the development of inexpensive input and output devices—computers, telephones, facsimile machines, and so on—that are affordable and easy to use. But the opposite is happening in the development of advanced computers and communications technology. High-performance computers, data networks, and archiving facilities are extremely expensive to build and operate. They are complicated technologies that require experts to develop and operate. Such expensive and complicated systems must be located in a few central facilities and made available to network users through such installations as the National Supercomputer Centers, NSFNET, the planned Human Genome Data Archive, and so on.

Allocating Resources and Assuring Equitable Access

The network raises a number of allocation and access issues that must be resolved. The number of high-performance computers and elaborate research instruments, such as telescopes or particle accelerators, are limited because of their high capital and operating costs. Universal access is not feasible, yet these facilities are critical to certain types of research. An equitable, fair process for allocating time on these facilities is crucial.

Network utility features, such as electronic mail, bulletin boards, journals, and so on, are basic to research in any field. Without them, one is locked out of the profession of science. Every researcher must have access to these services.

Updating Information Policies

Policies that currently govern the existing networks were developed to resolve conflicts over access and control of the information, e.g., protecting the privacy and confidentiality of communications and data on the network, or enforcing intellectual property rights. There are many more information policy questions concerning the rights and responsibilities to various electronic forms of communication that must still be addressed. Should "electronic mail" be protected like first class mail? Should bulletin board operators be legally responsible for messages placed on their boards? Does the First Amendment of the U.S. Constitution protect the sender? Should intellectual property protections be granted to electronic databases? If so, what form? Answers to these and other information policy questions will determine how the network is used and what services will be offered.

Adapting Science Policy

Just as an information infrastructure may change the way science is *done*, it may also lead to the need to change Federal science policy to accommodate these changes. High-performance computers may

[10] OSTP defines Grand Challenges as ". . .a fundamental problem in science or engineering, with potentially broad economic, political, and/or scientific impact, that could be advanced by applying high performance computing resources." Examples include: 1) Computational fluid dynamics for the design of hypersonic aircraft or efficient automobile bodies and recovery of oil; 2) Computer based weather and climate forecasts, and understanding of global environmental changes; 3) Electronic structure calculations for the design of new materials such as chemical catalysts, immunological agents and superconductors; 4) Plasma dynamics for fusion energy technology and for safe and efficient military technology; 5) Calculations to improve the understanding of the fundamental nature of matter, including quantum chromodynamics and condensed matter theory; and 6) Machine vision to enable real-time analysis of complex images for control of mechanical systems. *See* Office of Science and Technology Policy, *The Federal High Performance Computing Program* (Washington, DC: 1989), p. 8.

change research priorities and create new ways for research groups to organize and work together.

Determining the Type of Technology

The purpose of the technology will drive future policies. Management questions arising from the design and operation of a nationwide, ultra-high speed communication network may differ in nature from the problems of supporting and operating National Supercomputer Centers. But decisions related to both will collectively determine how effectively a national information infrastructure will be used in research and education.

On the other hand, information policy issues—that relate to the information that flows through the network-connected facilities—are seamlessly linked. Although the technical means for protecting and controlling information moving over the system may differ from computer to computer, or application to application, information policies are less dependent on the nature of the technology than on the generic issues. Privacy protection, access control, data security, and intellectual property protection are problems that need to be addressed across the board. Similarly, changes in the framework of Federal Government support and oversight of science policy will affect all technologies, disciplines, and agencies.

Determining Access

Depending on how one views the NREN, it is seen serving widely different user groups, ranging from a few federally funded high-end researchers engaged in "Big Science," to the scholarly community, to education from kindergarten through secondary schools. Both technical design decisions and policy will affect these various users in different ways. Who the intended user will be is a critically important consideration in making NREN policy. It is a subject dealt with in detail in a forthcoming OTA Report that focuses on the network as a broadband advanced communication infrastructure to simultaneously deliver data, video, and voice service.

Major Strategic Concerns

The mutual dependence and interconnectedness of a national information infrastructure will force the Federal Government to develop long-term strategies to guide the overall development of the NREN: this [illegible] be done in concert with a coordinated program to provide high-performance computer-based tools for science, research and education.

Breadth of Scope

Long-Range Planning Needs

Creating the infrastructure, the network, and its related resources, is not a one-time job. There are misconceptions that information infrastructure is a static concept only needing to be plugged in. This is not so; new applications will appear and the capabilities of technology will continue to grow and change. The system will, therefore, be a continually evolving assembly of technologies and services. Therefore planning and operating the NREN must be considered a dynamic process. An institutional framework must be developed to ensure its success.

Studies on information technology and science (including OTA's) rely on anecdotal examples, "gee-whiz" speculation about future applications, and the subjective views of the research and education community. These arguments are persuasive and sufficient to justify the support for the NREN, but they do not contribute sufficiently to long-term management and planning for the operation of the infrastructure. The Federal investment in computer, communication, and data resources for science and engineering should be based on a periodic assessment of needs and changing technologies. This assessment should include:

- surveys of existing resources—public, private nonprofit, and commercial, such as:
  —specialized and general purpose high-performance computing facilities;
  —Federal, State, and local data communication networks;
  —scientific and technical databases; and
  —software packages for research uses.
- utilization levels of existing facilities by categories such as:
  —field of research;
  —government, academic, or industrial use; and
  —research, graduate education, undergraduate education, or pre-college education.
- Barriers to efficient use of facilities, such as:
  —policy or legal barriers;
  —lack of standards for interconnection of systems; and
  —user difficulties such as lack of training, inadequate user interfaces, or lack of software and services.

- Projections of future computing needs (particularly, assessments of the need for new, large-scale research initiatives.)

Although NSF attempts to keep tabs on the computational and information needs of the science community, the pace of technological development and massive new science projects make such information more important during periods of tight budgets. These data are difficult to compile, and special efforts are needed to provide such planning data to those decisionmakers responsible for anticipating future national computing needs.

Chapter 2
Policy Considerations for High-Performance Computing

Currently the National Science Foundation (NSF) sponsors five leading edge computational centers, the four national supercomputer centers and the National Center for Atmospheric Researcher (NCAR) (see app. A). When the centers were established, one goal of the NSF initiative was to nationally provide researchers with access to leading edge technology. Prior to the NSF program, U.S. researchers and scientists had little opportunity—outside of Federal laboratories—to access supercomputers. Since their creation, the centers have been extremely successful in providing access to supercomputing resources to academic and industrial researchers.

The success of the NSF centers has made them the target of a debate over funding strategies for their support. It is noteworthy that they are not the only such facilities funded by the Federal Government or even by NSF. Computers, especially large-scale computers, always have required relatively large institutional structures to operate. The Department of Energy (DOE) and the Department of Defense (DoD) fund many more computational centers at a considerably higher cost than the NSF. Government establishment and support of scientific computing facilities date back to the earliest days of computing. Furthermore, high-performance computing is becoming increasingly important to all of science and engineering. The issue is not whether science, education, and engineering in the United States need high-performance computing centers, but rather how these centers should be supported, and how the costs of that support should be allocated over the long term.

It is imperative that the United States: 1) continue to steadily advance the capabilities of leading edge computer technology; 2) provide the R&D community with adequate computing resources; and 3) expand and improve the use of high-performance computing in science and engineering.

Advancing Computer Technology

Computers lie on the nearly seamless lines between basic research, applied research, and the development of new technologies. A program intended to advance the state-of-the-art of high-performance computing must include:

- physics research on fundamental devices, superconductors, quantum semiconductors, optical switches, and other advanced components;
- basic research in computer science and computer engineering, including theoretical and experimental work in computer architecture and a variety of other fields such as distributed systems, software engineering, computational complexity, data structures, programming languages, and intelligent systems;
- applied research and assembly of experimental laboratory testbed machines for exploring new concepts;
- experimentation, evaluation, and development of software for new prototype computers, e.g., the Connection Machine, Hypercube or neural nets;
- development of human resources and facilities for computing research needed to support a high-performance computing initiative, which requires additional trained researchers and research facilities;
- research and development of new technologies for data storage and retrieval (this may be the biggest technological bottleneck in the future); and
- creation of new algorithms tailored for advanced architectures to meet the needs of scientists and engineers for greater computational capabilities.

Difficulties and Barriers

Funding

The term "computational science" is used to define research devoted to applying computers to computationally intensive research problems in science and engineering. It is focused on developing techniques for using high-performance computing to solve scientific problems in fields such as chemistry, physics, biology, and engineering. Though growing, the base funding level for computer and computational science and engineering is currently low.

Defense Advanced Research Projects Agency (DARPA), NSF, DOE (particularly national laboratories such as Los Alamos and Lawrence Livermore Laboratory), the National Institute of Science and Technology (NIST), National Aeronautics and Space

Administration (NASA), and the National Institutes of Health have all contributed to improving the state-of-the-art of computer technology and its application to science and technology. There is, however, no clear lead agency to focus a national high-performance computing program.

A significant or substantial increase in the support of computational science as part of a high-performance computing initiative would require a relatively large additional investment. There is disagreement among researchers in the various disciplines about increasing funding for computational science. Some fear that investments in this area would reduce funds availk.ble for other research activities.

Procurement Regulations

In addition to the expense involved, obtaining prototype machines for experimental use has become more difficult because of some agency interpretations of Federal procurement law. In the past, research agencies have stimulated the development of advanced computer systems by purchasing early models for research use. Contracts for these machines were sometimes written before the machine was manufactured. The agency would then participate in the design and contribute expertise for software development. This cooperative approach was one key to advancing high-performance computing in the 1960s and 1970s. Unfortunately, the process has become more difficult as Federal procurement regulations for computing systems have become tighter and more complex.

Policy Issues

Federal support of computing R&D is intertwined with the political debates over technology policy, industrial policy, and the appropriate balance of responsibility between the Federal Government and the private sector in developing computer technology. Computing researchers study basic, and often abstract, c ncepts including the nature of complex processes and algorithms. But, the results of their work can have important practical implications for the design of computer hardware and software.

Computing research is often based on the study of prototypes and artifacts rather than natural phenomena. Consequently, Federal support is sometimes viewed as technological—rather than scientific—in nature. Moreover, Federal defense procurement directly supports the U.S. computer and software industry. Because of this relationship with industry, the High Performance Computing Initiative : .variably blends the role of traditional Federal science policy with Federal efforts to support precompetitive activities of a strategically important industry. This has led to confusion and debate over the goals and appropriateness of the proposed High Performance Cumputing Initiative.

Providing Access to Resources

Federal support for educational and research computer resources must broker their use among many different users with different needs at many different institutions. Policies that serve some users well may shortchange others. There are three general objectives that serve all: 1) provide funds for acquiring computer hardware and software; 2) assist in meeting operational expenses to maintain and manage facilities; and 3) ensure that scarce computational resources are distributed fairly to the widest range of users.

No single Federal program for supporting scientific computing is likely to serve the needs and policy objectives for all facilities and user groups. Support must come from a variety of coordinated programs. For example, since the inception of NSF's Advanced Scientific Computing programs, debates over support of the national supercomputer centers have reflected many different, and often contradictory, views of the roles the centers should play and the constituencies they serve.

Difficulties and Barriers

Diversity of Sources

Computers are expensive to buy and to operate. For larger machines, usage crosses many disciplines and users are associated with many different acad 'c institutions and industrial organizations. Supplying computer time can be a significant burden on research budgets, and support is often found by pooling funds from several sources.

No Natural Limits

Researchers seem to have an insatiable appetite for computer time. This perplexes policymakers who are used to dealing with expenditures for fixed cost items. One can estimate the number and kind of laboratory apparatus a chemist might need or microscopes a biology laboratory can use, based on the physical requirements of the researchers. However, the modeling of a complex organic chemical molecule for the design of a new pharmaceutical could saturate significant supercomputer resources. The potential use for supercomputer capacity appears to be limitless.

Administrators at research laboratories and government funding agencies have difficulties assessing computing needs and justifying new expenditures, either for purchase of additional computer time or for investments in upgrading equipment. It is even harder to predict future needs as researchers conceive new applications and become more sophisticated in developing innovative computer uses. These conflicting demands on the Federal science budget require careful balancing.

Disincentives to Investment

Support for computing resources may come from individual institutions themselves by underwriting the capital investment. The capital investment and operation costs are partially recaptured through fees charged back to the users. However, this model has not worked successfully, for a couple of reasons.

First, a multimillion dollar high-performance computer is a risky investment for an individual research institution. The risk is even greater for experimental machines whose potential use is difficult to anticipate. The institution must gamble that: 1) there is sufficient potential demand among research staff for the facilities; 2) federally supported researchers will have adequate funds to cover the costs; and 3) researchers with funds will choose to use the new computer rather than an outside facility.

Networks expand the possible user community of the facility, but they also provide access to competing systems at other institutions. In the past, researchers were, by and large, captive users of their own institutional facilities. Networks free them from this bondage. Now, researchers can use "distributed" computer resources elsewhere on the network. Faced with a wider "market" for computer time, research institutions may have less incentive to invest in more advanced systems, and instead upgrade local area networks to link with the NSFNET high-capacity backbone. On the other hand, networks can improve the efficiency and cost-effectiveness of computing by distributing computing capabilities.

Pricing policies for computer time must be carefully scaled to recover the costs of capital investments in hardware. High-computing costs can result in loss of revenue as researchers seek better rates at other institutions. The government requires that federally supported researchers pay no more than nonsupported researchers for computer time. But to ensure that operations break even, computer centers are forced to charge a rate equal to the costs divided by usage. This policy seems reasonable and equitable on the surface, but it results in higher rates for computer time when machine usage is light and lower rates as it grows. This pattern produces an upside-down market similar to that of the electric utilities before capital costs forced them to shave peak loads by charging a premium for power during periods of high usage. This is the reverse of airline rates where fares are lower when seats are empty and higher when planes are full.

Support Strategies

These disincentives and barriers have tended to limit investments in high-performance computers for research at a time when an increasing amount of important research requires access to more computational capacity. The Office of Science and Technology Policy's (OSTP) High Performance Computing Initiative, funding agencies' program plans, and pending legislation are aimed at balancing the Nation's R&D needs with high-performance computing capacity.

Four basic funding strategies to achieve this goal are described below:

Fully Support Federally Owned and Operated Centers

The most expedient strategy is to establish government-owned and operated facilities. The government could directly fund investments for hardware and software, and the centers' operational costs. There currently are several government funded and operated computational centers administered by the mission agencies. (See app. A, table A-1.) Government-financed computational centers provide a testbed for prototype machines and novel architectures that can help bolster the U.S. computer industry against foreign competition. Software development, critically needed for high-performance computing, is commonly a major activity at these centers.

A Federal high-performance computing initiative could select specific computational centers for full funding and operation by the Federal Government. A Federal agency might be needed to supervise the creation and management of the centers. Hardware would be owned or leased by the government. The center might be operated by a government contractor. The personnel, support staff and services, could either work directly for the government or a government contractor. These centers would be in addition to the existing mission agency computing centers.

Federally owned and operated computational centers currently exist under the management of several Federal mission agencies. The national laboratories—Los Alamos, Sandia, and Livermore—are operated by the DOE. Much of their work relates to national security programs, such as weapons research. NASA, DoD, and the Department of Commerce operate high-performance computing centers. NASA's centers primarily conduct aerospace and aerodynamic research. DoD operates over 15 supercomputers, whose research ranges from usage by the Army Corps of Engineers to Navy ship R&D to Air Force global weather prediction to intelligence activities of the National Security Agency (NSA). However, they do not fill the general needs of the science and education community. Access to these mission agency centers is limited, and only a small portion of the science community can use their facilities. The Federal Government could similarly own and operate computational centers for academic missions as well.

While federally owned and operated computing centers might risk experimentation with novel, untested computer concepts that academic or industrial organizations cannot afford, there is a possibility that this strategy could blossom into an additional layer of bureaucracy. The advantages of having direct government control over allocating computer time based on national priorities and acquiring leading edge technologies is offset by the risk of having government managers making decisions that should best be made by practicing scientists and engineers as is currently done at the NSF centers. Such shortcoming in systems management may be overcome by using nongovernment advisors or boards of governors, but centers could find it difficult to ensure stable year-to-year funding as national budgets tighten and competition for research dollars increases.

Fully or Partially Support Consortia or Institutionally Operated Centers

Federal science agencies can provide partial or full support to institutions for purchasing new computers. This is currently done by NSF and DOE. NSF provides major funding for four national supercomputer centers and the National Center for Atmospheric Research (NCAR) facility. DOE partially funds a supercomputer facility at Florida State University. The agencies provide funds for the purchase or leasing of computers and also contribute to the maintenance of the centers and their support staff. This has enabled the centers to maintain an experienced staff, develop applications software, acquire leading edge hardware, and attract computational scientists.

The government, through the NSF, provided seed funds and support to establish the centers and operate them. The NSF centers are complete computional laboratories providing researchers with leading edge technology, support services, software development, and computer R&D. The States and institutions in which the NSF centers are located have contributed about 35 percent of the expenses of the centers, and in addition the private sector has also contributed to the centers through direct funding and with in-kind contributions. Private firms are able to become partners with and use the centers' resources in return for their contribution. The national centers have attracted a user base exceeding that of the mission agency computational centers and including nearly every aspect of research, science, and education in U.S. universities.

The allocation of resources at these centers differs from that of the mission agency centers. The process of obtaining computing time at these centers is more open and competitive than at government-operated centers. The competitive process is aimed at fair allocation of the computing resources through a peer review process. Government subsidization of the operation of the computing centers has increased the use of computational resources, and increased the user base. For example, before the NSF national centers, there were only three or four places in the United States where high-performance computers were available if the research was not funded by mission agencies. Now, a growing number of States and universities operate computational centers to support research.

Some individuals have proposed that certain high-performance computing centers be assigned specialized missions. For instance, one center might emphasize biomedical research, or fluid dynamics; another, the responsibility for one of the other "grand challenges," such as global warming. NCAR is often used as an example of a successful discipline-oriented computational center to be used as a model for further specialization.

NCAR's computational center is partially funded by the NSF, but its research is specific to its mission in atmospheric science. In this way, it differs from the other four national NSF centers. NCAR's research includes climate, atmospheric chemistry, solar and solar-terrestrial physics, and mesoscale and microscale meteorology. The center houses a ccre staff of researchers and support personnel, yet its computational tools and human resources are available to the international atmospheric research community. Computer networks enable researchers around the Nation to access NCAR's facilities. NCAR, through its staff, research, hardware, and networks, has become a focal point for atmospheric research.

The advantage of a subject or discipline-specific computational center is that it focuses expertise and concentrates efforts on selected, important national problems. The staff is familiar with the type of work done within the disciplines and often knows the best ways to solve specific problems using computational science. Computers can be matched to fill the specific needs of the center rather than attempt to use a general purpose machine to serve (sometimes inadequately) the needs of diverse users. Experts in the field would have a central focus for meeting, comparing and debating research findings, and planning future research strategies much as atmospheric scientists now do at NCAR.

There are also disadvantages to discipline-specific centers. The "general" high-performance computing centers are a focal point for bringing together diverse users and disciplines. Researchers, scientists, computer scientists and engineers, and software engineers and designers work collaboratively at these centers. This interdisciplinary atmosphere makes the centers a natural incubator for the advancement of computational science, which is an essential component of research, by fostering communication among experts in various fields. It is noteworthy that NCAR, a mission-specific center, has a general purpose supercomputer identical to that at the general high-performance computing centers (i.e., a Cray Y-MP). Moreover, many atmospheric scientists also compute at the other NSF supercomputing centers.

The NSF centers were established to foster research and educational activities so that academic research could keep up with the needs and progress of the Federal research laboratories, the U.S. industrial research and engineering community, and foreign competitors, but subsidizing a select group of centers may create an impression of "elitism" within the science and technology community. The current funding of NSF centers authorizes only four federally funded centers. There has been no open competition for other computational centers in the NSF process since the selection in 1983-84, so equity within the community is often questioned. But the centers' plans are reviewed annually, and a comprehensive review was undertaken in 1989-90 that culminated in the closure of the Princeton University center. Some nonfederally funded State and university centers question why these installations are perpetually entitled to government funds while others are closed out of the competition.[1]

NSF's subsidization of its centers tends to establish a hierarchy within the computational community. However, objective competition among the centers would be hard to referee since the measures for determining eminence in computation are imprecise and subjective at best. The government must be leery of creating proclaimed "leaders" in computational science, because it risks setting limits instead of pushing the frontiers of computing.

Provide Supercomputing Funds to Individual Research Projects and Investigators

The Federal Government could choose to support computational resources from the grass-roots user level instead of institutional grants. Federal science agencies could provide funds to researchers as part of their research grants to buy and pay for computer services. In this way, the government would indirectly support the operational costs of the centers. Capital improvement would likely still need support

[1] Gillespie, Folkner & Associates, Inc., "Access to High-Performance Computer Resources for Research," contractor report prepared for the Office of Technology Assessment, Apr. 12, 1990, p. 36.

from the Federal Government because of the unpredictability of funding through user control and the need for long-term planning for maintaining and upgrading computer technology.

Some believe that funding the researcher directly for purchasing computer services would create competition among computational centers that could lead to improvements in the efficiency of the operation of computer centers and make them more responsive to the needs of the users. If scientists could choose where to "purchase" supercomputing services, they would likely choose the center that provides the best value and customer service. Scientists could match the services they seek with the specialties of each center to meet their individual needs. Proponents of funding computer services through individual research grants believe that creating efficient, market-oriented computational centers should be a goal of the high-performance computing program.

Centers vying for users might be captured by the largest users since they would have the most computing funds to spend. Well-funded users could force centers to cater to their needs at the expense of smaller users by the sheer purchasing power they represent. The needs of small users and new users could be slighted as centers compete for the support from big users. Competition among centers for users could have a downside if it should lead to isolation and lack of cooperation, and interfere with communication among the centers.

Upgrades and new machines involve large financial investments that use.-derived funds may not be able to provide. The uncertainty of future funding in a competitive environment would make long-range planning difficult. High-performance computers generally must be upgraded about every 5 years because the technology becomes outdated and maintenance too costly. National centers aimed at maintaining leading edge technology must upgrade whenever state-of-the-art technology emerges. Therefore, supplemental funding would be required for capital outlays even if user funds were used to offset operational expenses.

Critics of direct funding of researchers for supercomputer time claim that the money set aside for supercomputing should be dedicated solely for that use. They believe that if researchers were given nonearmarked funds for computer services, they might use them instead to buy minisupercomputers or graphic workstations for themselves, or to fund graduate students. They believe that much of the money would never reach the supercomputing centers, leading to unstable and unpredictable budgets. Direct funding of researchers for computing time was tried in the 1970s, and led to many of the problems identified in the Lax report.

Proponents of user-controlled funding believe that researchers can best decide whether supercomputing is necessary or not for their projects, and if minisupercomputers would suffice, then perhaps that is the best option.

Provide Incentives for State/Private Institutions To Supply Computational Services

Universities are heavily investing in information technologies and computational resources for the sciences. These non-Federal efforts should be encouraged. The government could provide matching funds to State and private institutions to contribute to the capital costs for computers and startup. Even a small amount of government seed money can help institutions leverage funds needed to establish a computing center. Supplemental assistance may be needed periodically for upgrading and maintaining up-to-date technology.

Some believe that temporary financial seeding of new centers is the best way for the Federal Government to subsidize supercomputing. Providing matching funds for several years to allow time for a center to become self-sufficient may be the best strategy for the Federal Government to assist in achieving supercomputing excellence.

After the seed period expires, centers must eventually upgrade their machines. Without additional funds to purchase upgrades they might fall behind new centers that more recently purchased state-of-the-art technology. Should this happen, a number of computational centers might be created, but none of them may end up world-class centers.

Expanding and Improving Usage

High-performance computers are general analytical tools that must be programmed to solve specific computational problems. Learning how to use the potential power of high-performance computers to solve specific problems is a major research effort itself. Research on how to apply high-performance computers to problems goes hand-in-hand with research on how to design the computers themselves. A Federal program to advance high-performance computing must strike a careful balance by supporting programs that advance the design of high-performance computers while at the same time advancing the science and engineering of computing for the R&D community.

It is important to distinguish *computational science* from *computer science and engineering*. Computer science is the science in which the object of intellectual curiosity is the computer itself. Computational science is the science in which the computer is used to explore other objects of intellectual curiosity. The latter discipline includes fields of basic research aimed at problems raised in the study of the computer and computing. They are not driven by specific applications. Although distinct, the two fields are closely related; researchers in each area depend on results and questions raised in the other.

Broader applications of computers often flow from advances made in research computing. Research in visualization, driven by the need to better understand the output of scientific calculations, has led to computer graphics technology that has revolutionized the movie and television industry and has provided new tools for doctors, engineers, architects, and others that work with images.

To advance the science of using high-performance computing, Federal programs must support five basic objectives:

1. **Expand the capabilities of human resources**—Individuals educated, trained, or skilled in applying the power of high-performance computers to new problems in science and technology are in high demand. They are sought by businesses, industries, and an assortment of institutions for the skills they bring to solving complex problems. There is a shortage of scientists, engineers and technicians with such skills. A Federal high-performance computing initiative must ensure that the pipeline for delivering trained personnel remains full.
2. **Develop software and hardware resources and technologies**—The research and development of technologies that can be applied to major research problems—"grand challenges" —must continue. Special efforts are needed to ensure progress in the development of software in order to harness the power of high-performance computing for the solution of R&D problems.
3. **Strengthen the scientific underpinnings of computation**—This can be accomplished through the support of computer science and engineering as well as computational science.
4. **Construct a broadly accessible, high-speed advanced broadband network**—Such a network will provide the scientific and educational community with access to the facilities, the data, and the software needed to explore new applications.
5. **Develop new algorithms for computational science**—Algorithms are mathematical formulas used to instruct computers (part of computer programs and hardware). They are the basis for solving computational problems. New and better algorithms are needed to improve the performance of hardware and software in the computing environment.

Difficulties and Barriers

Computer and computational sciences compete with many other disciplines, for science funding. They are relatively young fields and are growing from a small funding base. Funding levels for computing research is relatively small compared with the more mature disciplines. Stimulating growth in computer and computational science encounters a "chicken and egg" problem.

The size and level of activity of a research field is partially related to funds available. A Federal initiative designed to increase the research activity in computer and computational sciences must anticipate additional demands for Federal research funds. Furthermore, to maintain a healthy level of research activity, adequate funds to ensure future growth must be provided or talent will abandon the field to seek research money elsewhere. The small number of researchers working in computer and computational science may be cited as justification for *not* increasing levels of support, yet low levels of support limit the number of researchers and research positions.

Computational science is, in all but a few disciplines, a relatively new field. New researchers looking to establish their careers need assurance that their work will be recognized and accepted by their peers. Peer acceptance affects both their ability to obtain research funds and to publish articles in scientific journals. If computational methods are new to the field, the researcher may face a battle to gain acceptance within the traditional, conservative disciplines.

In many cases, researchers are in the early stages of understanding how to program radically new types of computers, such as massively parallel computers and neural nets. Researchers wishing to use such a computer need the assistance of those who can program and operate these computers for the duration of a project. There is currently a scarcity of such talent.

A NSF program dedicated to computational science and engineering may be needed. The program could fund computational scientists from a cross section of traditional disciplines such as biology, chemistry, and physics. Funds for programs aimed at developing human resources, such as fellowships, young investigator grants, and so on, may also need to be earmarked for computational science. Direct funding for computational sciences would overcome the tendency of the disciplines to favor the funding of conventional research and their reluctance to try new methodologies.

Computational Centers

The most difficult issues, which programs in NSF's Advanced Scientific Computing Division are addressing, stem from the problems in putting leading edge technology in the hands of knowledgeable users who can explore and develop its potential.

In the mid-1980s, NSF formed five national supercomputer centers. Three of them—the University of California at San Diego, Pittsburgh, and University of Illinois at Champaign-Urbana—were based on Cray supercomputers. One, at Cornell University, installed modified IBM computers, and the Princeton Center was based on a machine to be built by ETA, a subsidiary of Control Data that has since gone out of business. Subsequently, NSF did not renew the Princeton Center for a second 5-year period.

There have been many changes in the high-performance computing environment since the establishment of those centers. These changes include: 1) the evolution of the mini-supercomputer, 2) the establishment of other State and institutional supercomputing centers, 3) the increase in use and interest in applications of high-performance computing to research, 4) the emergence of the Japanese as a force in the design, manufacturing, and use of high-performance computers, and 5) the emergence of a national network. Because of these changes—particularly in light of budget pressures and the high cost of the program—questions are being asked about the future directions of NSF support for these centers.

The basic conflict arises from several concerns:

1. the need for the NSF programs that support computational centers to determine what their ultimate goals should be in an environment where technological changes and user needs are constantly changing;
2. the need of computer centers and their researchers for stable, predictable, and long-term support in contrast to the reluctance of the government to establish permanent institutions that may make indefinite claims on Federal funding; and
3. the view that any distribution of NSF high-performance computing funds should be openly competitive and based on periodic peer review.

Purposes for Federal High-Performance Computing Programs

Leading Edge Facilities

Leading edge facilities provide supercomputing to academe and industry and provide facilities for testing and experimenting with new computers. Academics are provided an opportunity to train with leading edge technology; researchers and engineers learn about new computer technology.

A leading edge facility's responsibilities go beyond merely providing researchers access to CPUs (central processing units). Manufacturers of high-performance computers rely on these centers to test the limits of their equipment and contribute to the improvement of their machines. Leading edge technology, by its nature, is imperfect. Prototype machines and experimental architectures are provided a testbed at these centers. Scientists' experiences with the technology assist the manufacturers in perfecting new computing equipment. Bottlenecks, defects, and deficiencies are discovered through use at the centers. Moreover, user needs have led to the creation of new applications software, computer codes, and software tools for the computers. These needs have forced the centers to take the lead in software development.

Several computational centers have industrial programs with large corporate sponsors. These corporations benefit from leading edge computational centers in two ways. First, industry gains access to the basic research conducted at universities on supercomputers. Second, industry learns how to use leading edge computer technology. The support services of these facilities are available to corporate sponsors and are a major attraction for these corporations. Corporate researchers are trained and tutored by the centers' support staff, and work with experienced academic users. They gain a knowledge of supercomputing, and this experience is taken back to their corporations. Participating corporations often leave the programs when they gain sufficient knowledge to operate their own supercomputer centers.

A high-performance computing plan that establishes and maintains leading edge facilities benefits a broad range of national interests. Academics learn how to use the technology, manufacturers use their experiences to improve the technology, and industry gains an understanding of the value of supercomputing in the work place.

Increasing the Supply of Human Resources

An important aspect of any high-performance computing program is the development of human resources. National supercomputer centers can cultivate human resources in two ways. First, researchers and scientists are taught how to use high-performance computers, and new users and young scientists learn how to use modern scientific tools. Second, national centers provide an atmosphere for educating and cultivating future computer support personnel. Users, teachers, and technicians are critical to the future viability of supercomputing.

Producing proficient supercomputer users is an important goal of a high-performance computing program. Researchers with little or no experience must be trained in the use of the technologies. Education must begin at the graduate level, and work its way into undergraduate training. Bringing supercomputer usage into curricula will help familiarize students with these tools. The next generation of scientists, engineers, and researchers must become proficient with these machines to advance their careers. The need for competent users will increase as supercomputers proliferate into the industrial sector. Already there are reports of a shortage of supercomputer trained scientists and engineers.[2]

Support staff is an essential element of computational centers. The support services, which include seminars and consultation and support, educate the next generation of users. Support personnel are the trouble-shooters, locating and correcting problems, and optimizing computer codes. The NSF national centers have excellent staff, some of whom have moved to responsible positions at State and university-operated centers. The experience they gained at the NSF national centers contributes to the viability of new high-performance computing operations in industry and elsewhere in academe. The importance of the services that support personnel provide is often overlooked by policymakers, yet their contributions to supercomputing are invaluable. The greatest asset of a proficient high-performance computing center is the staff, not the computer. A high-performance computing program must emphasize the importance of developing human resources by producing educated users and users who will educate.

Advancing Computational Science

High-performance computer centers are a focal point for bringing together diverse users and disciplines. Researchers, scientists, computer scientists and engineers, and software engineers and designers work collaboratively at these centers. This interdisciplinary atmosphere makes the centers a natural incubator for the advancement of the computational sciences, which is an essential component of supercomputing. A national high-performance computing program could promote the computational sciences by fostering communication among experts in various fields.

Researchers and scientists know what questions to ask, but not necessarily how to instruct computers to answer them. Computational scientists know how to instruct computers. They create the computer instructions sets, computer codes, and algorithms for computers so that researchers can most efficiently utilize the technology. The development of computer codes and software is often a collaborative effort, supported by previous codes, software tools, and support staff, many of whom are computational scientists. Providing the methodology for utilizing

[2] Michael Schroeder, "How Supercomputers Can Be Super Savers," *Business Week*, Oct. 8, 1990, p. 140.

these tools is as important as providing the tools themselves.

Developing New Software Applications

New algorithms and codes must be developed to allow optimum use of supercomputer time. One of the more frequent criticisms of many high-performance computing operations has been the use of suboptimal codes. Supercomputer time is wasted when outdated or less than optimal codes are used. Creating codes is a specialty in itself. The development of codes is so labor intensive and time consuming that using an outdated code, as opposed to creating a new one, is sometimes more time efficient, although it may waste costly supercomputer time. A high-performance computing program could advance the usage of new and efficient codes by promoting computational science.

Providing Access to More Supercomputing CPUs

Supercomputing CPUs offer researchers computing power and speed unattainable from conventional mainframes. High-performance computer centers provide, at a minimum, access to supercomputing cycles. Supercomputing CPUs currently are a scarce resource in high demand. Any Federal high-performance computing program will increase the amount of supercomputing cycles available to researchers. It is uncertain, however, how much increase in CPUs the government should provide. Supercomputers are used in the advancement of all scientific disciplines, for both "big" and "little" science projects. All areas of research benefit from high-performance computing. Notwithstanding any reasonable level of effort, the government will be unable to provide enough supercomputing resources to meet all researchers' needs. They will always seek more and faster supercomputing power.

Computer facilities whose main goal is to provide supercomputing CPUs are often called "cycle shops." The NSF centers *are not* cycle shops. At cycle shops, support services are minimal: A skeletal support staff, enough personnel to keep the machines up and running, is all that is required. This limits cycle shops' usefulness to primarily experienced users. Only proven technology can be used. Training, education, and software development are not major activities at such facilities. User applications have to be "canned" and ready for use. These centers are the antithesis of leading edge facilities. Cycle shops are more economical for experienced users in need of large amounts of CPU time. This is not the majority of users, however.

Improving Data Storage Capabilities

Increasing importance is being placed on data storage capabilities. Researchers now realize the limits of current data storage technologies. A high-performance computing program can stimulate research in high-capacity storage and retrieval technologies.

Data storage technologies do not have the public appeal and visibility that supercomputers do. For this reason, they have been overlooked in supercomputing R&D, yet data storage is an integral part of high-performance computing. Supercomputers often use and produce large data sets. Computational centers are increasingly running into data memory and storage problems. New technologies for gathering data, e.g. satellites and automated sensors, are placing even greater demands on storage facilities. These data are often used in computing, and are converted into new data sets that require additional storage.

The Federal Government could take the initiative in R&D on new storage technologies, emphasizing its importance to high-performance computing. The amount of data handled at supercomputing centers will increase as the user base multiplies, and as data sharing increases through the use of high-capacity communications networks through the National Research and Education Network (NREN). Storage technologies are currently pushing their limits, and breakthroughs are needed if they are not to become the limiting factor in high-performance computing.

Chapter 3
High-Performance Computers: Technology and Challenges

Computers and the R&D Process

Scientists use the theories and techniques of mathematics for building and describing models in logical ways and for calculating the results they yield. As early as the third century B.C., the Alexandrian scholar Eratosthenes estimated the circumference of the earth to an accuracy within 5 percent of what we now consider to be the correct figure. He did so by making assumptions about the nature of the physical universe, making measurements, and calculating the results.[1] In essence, he did what modern scientists do. He constructed a hypothetical model that allowed him to apply mathematical tools—in this case, trigonometry and arithmetic—to data he collected.

Scientific models are used both to test new ideas about the physical universe and to explore results and conclusions based on those models. Eratosthenes discovered a new "fact"—the size of the earth. Had his calculations, instead, confirmed a result already discovered by some other means, he would have accomplished a different research purpose; he would have provided evidence that the model of the universe was correct. Had they differed with known fact, he would have had evidence that the model was incorrect. Science advances, step by step, through a process of building models, calculating results, comparing those results with what can be observed and, when observations differ, revising the models.

Modes of Research Computing

Just as mathematics is central to science, computers have become basic instruments of research to modern science and play a wide variety of roles. Each of the roles is based on mathematical modeling, using the interactive solution of thousands of equations.

To Perform Complex Calculations

Sometimes the basic mathematics and structure of a physical process are well known—the equations, that describe the flow of air around a solid object, for example. Researchers may wish to calculate the results of this process in experimental designs such as a new aircraft wing or the shape of an automobile. Calculating results from flow equations are enormously time-consuming even on the most powerful computers of today. Scientists must simplify these problems to fit the capabilities of the computers that are available. They sacrifice accuracy and detail in their model to achieve computability.

To Build New Theories and Models

At other times, researchers seek to understand the dynamics of a process, like the aging of a star or formation of a galaxy. They create computer models based on theories and observe how the behavior of those models do or do not correspond to their observations.

To Control Experimental Instruments and Analyze Data

Most modern scientific instruments have some computational power built in to control their performance and to process the measurements they make. For many of these, from the largest particle accelerators or space platforms to more modest instruments, the computer has become an integral and indispensable part.

Such research instruments generate enormous flows of information—some at rates up to several trillion units (terabits) a day. Unpackaging the data flow, identifying the elements, and organizing those data for use by scientists is, itself, a sizable computational task. After the initial steps, still more computer power is needed to search this mountain of data for significant patterns and analyze their meanings.

To Better Understand and Interact With Computer Results

At the most basic level, computers produce numbers; but numbers usually represent a physical object or phenomenon—the position of an atom in a protein molecule, the moisture content in a cloud, the stress in an automobile frame, or the behavior of an explosive. To make sense to researchers, the streams of numbers from a computer must be

[1] Thomas S. Kuhn, *The Copernican Revolution* (Cambridge, MA: Harvard University Press, 1985), p. 274.

converted to visual displays that are easier to understand when seen by the eye. Researchers are now concentrating on visualization—pictorial displays that incorporate images, motion, color, and surface texture to depict characteristics of an analysis on a computer screen.

Some researchers are exploring more advanced techniques that use other senses such as sound and touch to convey results to the human mind. By incorporating all of these technologies, they may eventually be able to create what is called "virtual reality," in which a scientist equipped with the proper gear could interact directly with a model as though he or she were standing in the midst of the phenomenon that was modeled. A biochemist could "walk" around and about a protein molecule, for example, and move atoms here and there, or a geologist could explore the inside of an active volcano.

To Provide "Intelligent" Assistance

Computer operations are not restricted to only computational operations on numbers. The popularity of word processors shows that computers can manipulate and perform logical operations on symbols, whether they represent numbers or not. Experts in the "artificial intelligence" community have been exploring how computers can assist researchers in ways other than direct computation of results. They have worked on systems that can prove mathematical theorems or perform tedious manipulations of algebraic expressions, systems that help chemists find new forms of molecules, and natural language inquiry systems for databases.

A national research and educational network (NREN) would create a critical need for such help in the future so that scientists are not overwhelmed by the complexity and amount of information available to them. New tools such as "knowbots"—small autonomous programs that would search databases throughout the network for information needed by the researcher—have been proposed.

Implications for Federal Programs

The traditional view of the "scientific computer" as one specifically intended for high-speed arithmetic computation is changing as researchers use computers for an increasingly rich variety of tasks. Any Federal initiative supporting computational science must create an environment that supports a wide variety of machines with improved capabilities, many of which serve specialized user communities.

Numerical computation is still critically important, but so are applications such as database manipulation, artificial intelligence, image production, and on-line control of experimental instruments. Even the design of computers meant to do numerical calculations is becoming more specialized to address specific types of problems.

The NREN is a crucial element of efforts to make high-performance computing widely available to the U.S. research community. Members of research groups who need these specialized computers are widely scattered throughout the country, and so are the computers they need.

The Evolution of Computer Technology

Government and Computer R&D

Like much of the new electronics technology of the day, computers in large measure grew out of work done during World War II for defense research programs. After the war, many engineers and scientists who staffed those programs took their knowledge into the private sector to begin the commercial U.S. computer industry.

The Federal Government remains a major purchaser, user, and force in shaping computer technology. Its influence is particularly strong in scientific computing; many computational researchers either work for the government in national laboratories or are substantially funded by government agencies. The computing needs of the defense agencies, and the weapons programs of the Department of Energy (earlier the Atomic Energy Commission (AEC)), demanded continual advancement of the speed and power of scientific computing.

Computers that meet the specifications of scientific users were not, until recently, commercially successful or widely available. As a result, Federal agencies needing these large scientific machines had to fund their development. Control Data's 6600 computer in the mid-1960s was among the first large scientific machines designed for national defense needs to be marketed successfully in the private sector.

Even though scientific computers were not originally successful in the nongovernment market, their technology was. The "Stretch" computer, designed and built by IBM for the AEC, provided many innovations that were later used in the design of the IBM 360 series that was the basic IBM product line for over a decade. Federal science agencies such as the National Science Foundation (NSF), Defense Advanced Research Projects Agency (DARPA), and the Office of Naval Research (ONR) have also contributed over the years to the development of computer architecture through their computer science and engineering research programs.

The government role in support of basic and applied research in computing and in testing prototype machines and making them available to researchers is critical to the well-being of small specialized firms in high-performance computing.

Government support for research in computer architecture has gone through cycles. In the early days, it was in research laboratories that computer scientists first developed many of the architectural concepts that formed the basis for general purpose computers. As computers became more complex and their manufacture a more refined art, academic research on computer design waned. Perhaps the decreased interest in architecture research resulted from the notion at that time that the major computer design issues had been settled and the development of new generations of machines should be left to the industry. The academic research that continued was mostly paper-and-pencil design simulated on conventional computers.

During the last decade, advances in microelectronics created opportunities to explore radical new designs with relatively inexpensive off-the-shelf chips from manufacturers, or custom designs. Experts were predicting the end of performance improvements that could be wrung from traditional design concepts, while the costs for coaxing performance improvements were increasing dramatically. As a result, computer scientists and engineers are again exploring alternate approaches, and academic research has now returned to the development and testing of prototypes, this time in cooperation with industry. Now, as then, the basic question is whether these experimental designs are more efficient and effective for performing specific types of calculations.

Computer scientists and engineers basically look in three directions to improve the efficiency and increase the speed of computers:

1. the fundamental technology of the computer components;
2. the architecture of the computer; and
3. the software programs and algorithms to instruct and control the computers.

These three areas of investigation are distinct fields of research, but they have an important influence on each other. New devices allow computer designers to consider different approaches to building computers, which, in turn, can lead to new ways of programming them. Influences can just as easily go the other way: new software techniques can suggest new machine architectures. One of the problems with introducing radically new types of computers into common use is that entirely new theories of programming must be developed for them, whereas software techniques for traditional machines have taken place over 40 or 50 years of development and refinement.

Fundamental Technologies

Basically, computers are complex assemblies of large numbers of essentially similar building blocks. These building blocks—all of which are generally different types of logical switches that can be set in one of two states (on-off)—are combined to form the memory, registers, arithmetic units, and control elements of modern digital computers (*see* box C). The advance of computer technology at this level can be seen as the clustering of more and more of these basic switches into increasingly smaller, faster, cheaper, and more reliable packages.

Integrated Circuits—Electrical engineers predict that, by 2000, chip manufacturers will be able to put over one billion logic gates (switches) on a single chip. Some silicon chips already contain more than a million gates. This level of complexity begins to allow producers to put huge computational power on one processor chip. By the end of the decade, it is expected that a single chip will have the complexity and the power of a modern supercomputer, along with a significant amount of memory.

This trend is influencing research in computer design. Computer scientists and engineers use the term "architecture" to describe the art of arranging the flows of data and the detailed logical processes within the computers they design. Given the com-

Box C—The Building Blocks of Modern Computer Hardware

From electro-mechanical relays to vacuum tubes to silicon-based very-large-scale integrated circuits, the electronic technologies that form the basic components of computers have steadily and rapidly advanced year by year since the 1940s. One measure of improvement is the number of transistors (the basic building block of logic and memory) that can be placed on a chip. Increase in transistor density is expected to continue throughout the coming decade, although "traditional" silicon technology, the basis of microelectronics for the last few decades may begin reaching its maximum cost/performance benefit. It may become too costly to derive future performance advancements out of silicon.

In the past, as each type of technology—mechanical switches, vacuum tubes, and transistors—reached its limits, a new technology has come along that allowed information technology to continue improving; this phenomenon is likely to continue. Researchers are exploring several basic technologies that, if successful, could continue these rates of growth, not only through this decade, but well into the next century.[1]

Gallium Arsenide Compounds

Gallium Arsenide (GaAs) is a compound with semiconductor properties similar to, but in some ways superior to, silicon. Spurred in part by interest from the Department of Defense, researchers have developed GaAs to the point where such devices are being produced for commercial application. But will it ever be cost-effective to manufacture devices complex enough and in quantities sufficient to build full-scale computers in a cost-effective way? Some manufacturers are trying.

Cray Computer Corp. (CCC), a separate company spun off from its parent Cray Research, and Convex Computers—a manufacturer of entry-level supercomputers—are attempting to use GaAs-based components for their new machines. Although offering much greater speeds for the machine, these components have proved to be difficult to manufacture and to assemble into a large-scale mainframe. Their efforts are being watched closely. Some experts think that some of these manufacturing difficulties are inherent and that GaAs will remain a valuable but expensive "niche" technology, possibly useful for high-speed and costly applications, but not serving as the "workhorse" all-purpose replacement for silicon in everyday applications.[2]

Superconductivity

For years it has been known that some materials attain a state known as "superconductivity" when cooled sufficiently. A superconductive material essentially transmits electricity without (or with low) resistance. Using superconductivity, a switch known as a "Josephson Junction" (JJ) can be built that could, in theory, serve as the basis of computer logic and memory.

The problem has been that "sufficiently cooled" has meant very cold indeed, nearly the temperature of liquid helium, only 4 degrees Kelvin.[3] Although it is possible to attain these temperatures, it requires extensive and complex apparatus either for refrigerating or for using liquid helium, a very temperamental substance to deal with. Problems with reliably manufacturing JJs have also been difficult to solve. Because JJs could move computer capabilities beyond silicon limits if these problems were solved, some manufacturers, particularly the Japanese, have continued to explore low-temperature superconductivity.

Within the last few years, however, the discovery of materials that exhibit superconductivity at higher temperatures has led to a renewed interest in the JJ.[4] "High temperature" is still very cold by normal standards, around 50 to 100 degrees Kelvin, but it is a temperature that is much more economical to maintain. Significant materials problems still confound attempts to manufacture JJs reliably and in the bulk necessary to manufacture computers. However, investigators have just begun exploring this technology, and many of them expect that these

[1] U.S. Congress, Office of Technology Assessment, *Microelectronics Research and Development—Background Paper*, OTA-BP-CIT-40 (Washington, DC: U.S. Government Printing Office, March 1986).

[2] Marc H. Brodsky, "Progress in Gallium Arsenide Semiconductors," *Scientific American*, February 1990, pp. 68-75.

[3] Kelvin is a unit of measurement that uses as its reference, "absolute zero," the coldest temperature that matter can theoretically attain. In comparison, zero degrees Centigrade, the temperature at which water freezes, is a warm 273 degrees Kelvin.

[4] U.S. Congress, Office of Technology Assessment, *Commercializing High-Temperature Superconductivity*, OTA-ITE-388 (Washington, DC: U.S. Government Printing Office, August 1988).

problems will be solved, in part because of the potential importance of the technology if it can be tamed. It has been suggested that Japanese manufacturers continue to work on low-temperature prototypes in order to gain experience in designing and building JJ-based computers that could be useful if and when high-temperature technology becomes available.

Other Advanced Technologies

Researchers are also investigating other promising technologies, such as "optical switching" devices. Fiber optics already offers significant advantages as a communication medium, but signals must be converted back to electrical form before they can be manipulated. It might be attractive in terms of speed and economy if one could handle them directly in the form of light.

Other researchers are working on so-called "**quantum effect**" devices. These devices use silicon—and in some cases GaAs—materials, but take advantage of the quantum, or wave-like, behavior of electrons when they are confined in very small areas (say, on the order of 100 atoms in diameter.)[5] Again, problems of manufacturing, particularly devices as small as this, present major difficulties to be overcome.

[5] Henry I. Smith and Dimitra A. Antoniadis, "Seeking a Radically New Electronics," *Technology Review*, April 1990, pp. 27-39.

plexity that modern chips can embody, a chip designer can use them to build bigger, more elaborate constructs. Such a designer might be thought of more as a "city planner"—someone who arranges the relationships between much larger structures and plans the traffic flow among them.

Computer design is helped considerably by modern technology. First, through use of automated design and "chip foundries" for producing customized chips (some of which can be accessed via a network), designers can move from paper-and-pencil concepts to prototype hardware more easily. Many of the new high-performance computers on the market use processor chips custom-designed for that specific machine; automated chip design and manufacture shorten the time and improve the flexibility in producing custom chips.

Second, the market offers a variety of inexpensive, off-the-shelf chips that can be assembled to create new and interesting experimental designs. One of the best known successful examples of this type of research is a project initiated at the California Institute of Technology. There, researchers designed and built a customized computer to help them with certain specialized physics calculations. They developed the first "hypercube" machine using a standard line of processor chips from Intel. Intel supported the project in the early days, principally through the donation of chips. Later, as the design concept proved itself and attracted the attention of government agencies, full-scale research support [illegible] provided to the group.

The impact of that low-budget project has been enormous. Several companies (including Intel) are in, or are planning to enter, the high-performance computer market with computers based on the hypercube design or one of its variations. Universities are beginning to realize the potential of specialized, low-budget machines, among them Caltech, Rice, and Syracuse. Three NSF centers (National Center for Supercomputing Applications, Pittsburgh Supercomputing Center, and the San Diego Supercomputer Center) also have installed these architectures for access by the nationwide academic community.

Based on the history and trends in computer architecture research, it appears that: 1) it is feasible to design and build computers with architectures customized for particular tasks; 2) the availability of powerful, inexpensive chips, has prompted academic laboratories to return to research in computer architecture; 3) new ideas in computer architecture can likely be commercialized quickly; and 4) universities that have access to fabrication facilities are more likely to develop new, specialized machines.

In the past, such customized machines would have been considered curiosities, with no chance of competing with traditional designs. The computer industry at that time was conservative, and users were unwilling to take chances on new ideas. Now, some entrepreneurs will gamble that if the system has distinct advantages in power and cost, new markets will open, even for systems based on radical new design theories.

But bringing a new high-performance machine to market is neither cheap nor simple. Millions of dollars—sometimes hundreds of millions—must be spent refining the design, developing software, and solving manufacturing problems, before a design concept moves from the laboratory into general use. The speed and ease of this transfer depends heavily on whether the technology is evolutionary or revolutionary.

It is difficult to say which computer technologies will become the foundation for building computers over the next decade. Despite the fact that all of the alternative technologies have difficulties to be overcome, it is likely that one or more new component technologies will be developed to fuel the rapid growth of computer capability into the next decade and beyond. But advances in fundamental technology alone will not be sufficient to achieve the increases in computer power that are needed by research users.

Computer Architecture

The term "computer architecture" denotes the structural design of a computer system. It includes the logical behavior of major components of the computer, the instructions it executes, and how the information flows through and among those components. A principal goal of computer architecture is to design machines that are faster and more efficient for specific tasks.

"Supercomputer" is commonly used by the popular media to describe certain types of computer architectures that are, in some sense, the most powerful available. It is not, however, a useful term for policy purposes. First, the definition of computer "power" is inexact and depends on many factors, including processor speed and memory size. Second, there is no clear lower boundary of "supercomputer power." IBM 3090 computers come in a wide range of configurations, but are they "supercomputers"? Finally, technology is changing rapidly, and with it the conceptions of the power and capability of various computers. Here, the term "high-performance computers" (HPC) (distinguished from the Federal program to advance high-performance computing referred to as the "high-performance computing initiative") includes a variety of machine types.

One class of high-performance computing consists of large, advanced, expensive, powerful machines, designed principally to address massive computational science problems. These computers are the ones often referred to as "supercomputers." Their performance is based on central processing unit (CPU) power and memory size. They use the largest, fastest, most costly memories. A leading edge "supercomputer" can cost up to $20 million or more.

A large-scale computer's power comes from a combination of very high-speed electronic components and specialized architecture. Most machines use a combination of "vector processing" and "parallel processing" (parallelism) in their design. A vector processor is an arithmetic unit of the computer that produces a series of similar calculations in an overlapping, assembly-line fashion (many scientific calculations can be set up in this way).

Parallel processing is the use of several processors that simultaneously solve portions of a problem that can be broken into independent pieces for computing on separate processors. Currently, large, mainframe high-performance computers such as those of Cray and IBM are moderately parallel, having from two to eight processors.[2] The trend is toward more parallel processors on these large systems. The main problem to date has been to figure out how problems can be set up to take advantage of the potential speed advantage of larger-scale parallelism.

The availability of software for supercomputer application is a major challenge for high-performance computing in general, but it is particularly troublesome in the case of large parallel processing systems. Parallel processing requires that the complexity of the problem be segregated into pieces that can run separately and independently on individual processors. This requires that programmers approach solutions in a very different manner from the way they program information flow and computations on vector processors. Until the art of parallel programming catches up with the speed and sophistication of hardware design, the considerable power of parallel computing will be underutilized. Software development for supercomputing must be given high priority in any high-performance computing initiative.

[2] To distinguish between this modest level and the larger scale parallelism found on some more experimental machines, some experts refer to this ited parallelism as "multiprocessing."

Some machines now on the market (called "mini-supers" or "minisupercomputers") are based on the structure and logic of a large supercomputer, but use cheaper, slower electronic components and lower performance technology. They are relatively less expensive than high-end supercomputers. These systems sacrifice some speed, but cost much less to manufacture. An application that is demanding but does not require a full-size supercomputer may be more efficiently run on a minisuper.

Other types of specialized systems also have appeared on the market. These machines gain computation speed by using fundamentally different architectures. They are known by colorful names such as "Hypercubes," "Connection Machines," "Data Flow Processors," "Butterfly Machines," "Neural Nets," or "Fuzzy Logic Computers." Although they differ in design concept, many of these systems are based on large-scale parallelism. Their designers get increased processing speed by linking large numbers—hundreds or even thousands—of simpler, slower, and cheaper processors. But computational mathematicians and scientists have not yet developed a good theoretical or experimental framework for understanding how to arrange applications to take full advantage of these massively parallel systems. Therefore, these systems are still, by and large, experimental, even though some are on the market and some users have developed applications software for them. Experimental as these systems are however, many experts believe that any significantly large increase in computational power must grow out of experimental systems such as these or from other forms of massively parallel architecture or hybrid architectures.

"Workstations," the descendants of personal desktop computers, are increasing in power; new chips being developed will soon offer computing power nearly equivalent to a Cray 1 supercomputer of the late 1970s. Thus, although high-end high-performance computers will oe correspondingly more powerful, scientists who wish to do heavy-duty computing will have a wide selection of options in the future. Policymakers must recognize that:

- The term "supercomputer" is a fluid one, potentially covering a wide variety of machine types; similarly, the "supercomputer industry" is increasingly difficult to identify as a distinct entity.
- Scientists need access to a wide range of high-performance computers from desktop workstations to full-scale supercomputers, and they need to move smoothly and seamlessly among these machines as their research needs require.
- Government policies should be flexible and broadly based to avoid focusing on a narrowly defined class of machines.

Mere computational power is not always the sole objective of designers. For example, in the case of desktop computers like the Apple MacIntosh or NEXT Computers, or the more powerful engineering workstations, much effort has gone into improving the communication between the machine and the operator (user interface). Computers are being designed to be more easily linked through data communication networks. Machines are being designed to do specialized tasks within computer networks, such as file management and internetwork communication. As computer designers develop a wider variety of machines specialized for particular tasks, the term "high performance" covers a wider range of applications and architectures, including machines that are oriented to numerical scientific calculation.

Computer Performance

Computers are often compared on the basis of computer power—usually equated to processing speed. The convention used for measuring computer power is "FLOPS" (floating point operations per second). The term "floating point" refers to a particular format for numbers (scientific notation) within the computer that is used for scientific calculation. A floating point "operation" refers to a single arithmetic step, such as multiplying or dividing two numbers, using the floating point format. Thus, FLOPS measure the speed of the arithmetic processor. Currently, the largest supercomputers have processing speeds ranging up to several billion FLOPS. DARPA has announced a goal of developing in this decade a "teraflop" machine, a computer that executes one trillion FLOPS.

Peak computer speed and computer systems performance are two different things. Peak computer speed is the raw theoretical performance that is the maximum possible for the computer architecture. Computer system performance, the actual speed under use, is always lower—sometimes much lower. Theoretical peak speed alone is not a useful measure of the relative power of computers. To understand why, consider the following analogy.

At a supermarket checkout counter, the calculation speed of the cash register does not, by itself, determine how fast customers can checkout. Checkout speed is also affected by the speed that the clerk can enter each purchase into the cash register and the time it takes to complete a transaction with each customer—bag the groceries, collect money, make change—and move on to the next. The length of time the customer must wait in line to reach the clerk may be the most important factor of all, and that depends on how many clerks and cash registers are provided.

Similarly, in a computer, how quickly calculations can be set up and input to the processor and how quickly new jobs and their data can be moved in, completed, and the results moved out of the computer determines how much of the processor's speed can actually be harnessed (some users refer to this as ''solution speed''). Solution speed is determined by a variety of architectural factors located throughout the computer system as well as the interplay between hardware and software. Similar to the store checkout, as a fast machine becomes busy, users may have to wait in line. From a user's perspective, then, a theoretically fast computer can still deliver solutions slowly.

To test a machine's speed, experts use ''benchmark programs,'' i.e., sample programs that reproduce a ''standard'' workload. Since workloads vary, there are several different benchmark programs, and they are continually being refined and revised. Measuring a supercomputer's speed is a complex and important area of research. Performance measurement provides information on what type of computer is best for particular applications; such measurements can also show where bottlenecks occur and, hence, where hardware and software improvements should be made.

One can draw some important implications from these observations on computing speed:

- Computer designers depend on feedback from users who are pushing their machines to the limit, because improvements in overall speed are closely linked to how the machines are programmed and used.
- There is no ''fastest'' machine. The speed of a high-performance computer depends on the skill of those that use and program it, and the type of jobs it performs.
- One should be skeptical of claims of peak speeds until machines have been tested by users for overall systems performance.
- Federal R&D programs for improving high-performance computing must stress software, algorithms, and computational mathematics as well as research on machine architecture.

Appendix A
Supercomputer Centers

The National Supercomputer Centers

In February 1985, National Science Foundation (NSF) selected four sites to establish national supercomputing centers: Cornell University, the University of Illinois at Urbana-Champaign, the University of California at San Diego, and the John von Neumann Center in Princeton. A fifth site, Pittsburgh, was added in early 1986. Funding for Princeton's Von Neumann Center was later dropped. The four remaining NSF centers are described briefly below.

The Cornell Theory Center

The Cornell Theory Center is located on the campus of Cornell University. Over 1,900 users from 125 institutions access the center. Although Cornell does not have a center-oriented network, 55 academic institutions are able to utilize the resources at Cornell through special nodes. A 14-member Corporate Research Institute works within the center in a variety of university-industry cost-sharing projects.

In November 1985 Cornell received a 3084 computer from IBM, which was upgraded to a four-processor 3090/400VF a year later. The 3090/400VF was replaced by a six-processor 3090/600E in May 1987. In October 1988 a second 3090/600E was added. The Cornell Center also operates several other smaller parallel systems, including an Intel iPCS/2, a Transtech NT 1000, and a Topologix T1000. Some 50 percent of the resources of Northeast Parallel Architecture Center, which include two Connection machines, an Encore, and an Alliant FX/80, are accessed by the Cornell facility.

Until October 1988, all IBM computers were "on loan" to Cornell for as long as Cornell retained its NSF funding. The second IBM 3090/600, procured in October, will be paid for by a NSF grant. Over the past 4 years, corporate support for the Cornell facility accounted for 48 percent of the operating costs. During those same years, NSF and New York State accounted for 37 percent and 5 percent, respectively, of the facility's budget. This funding has allowed the center to maintain a staff of about 100.

The National Center for Supercomputing Applications

The National Center for Supercomputing Applications (NCSA) is operated by the University of Illinois at Urbana-Champaign. The center has over 2,500 academic users from about 82 academic affiliates. Each affiliate receives a block grant of time on the Cray X-MP/48, training for the Cray, and help using the network to access the Cray.

The NCSA received a Cray X-MP/24 in October 1985. That machine was upgraded to a Cray X-MP/48 in 1987. In October 1988 a Cray-2s/4-128 was installed, giving the center two Cray machines. This computer is the only Cray-2 now at a NSF national center. The center also houses a Connection Machine 2, an Alliant FX/80 and FX/8, and over 30 graphics workstations.

In addition to NSF funding, NCSA has solicited industrial support. Amoco, Eastman Kodak, Eli Lilly, FMC Corp., Dow Chemicals, and Motorola have each contributed around $3 million over a 3-year period to the NCSA. In fiscal year 1989 corporate support amounted to 11 percent of NCSA's funding. About 32 percent of NCSA's budget came from NSF, while the State of Illinois and the University of Illinois accounted for the remaining 27 of the center's $21.5-million budget. The center has a full-time staff of 198.

Pittsburgh Supercomputing Center

The Pittsburgh Supercomputing Center (PSC) is run jointly by the University of Pittsburgh, Carnegie-Mellon University, and Westinghouse Electric Corp. More than 1,400 users from 44 States utilize the center. Twenty-seven universities are affiliated with PSC.

The center received a Cray X-MP/48 in March 1986. In December 1988 PSC became the first non-Federal laboratory to possess a Cray Y-MP. For a short time, both machines were being used simultaneously; however the center has now phased out the Cray X-MP. The center's graphics hardware includes a Pixar image computer, an Ardent Titan, and a Silicon Graphics IRIS workstation.

The operating projection at PSC for fiscal year 1990, a "typical year," has NSF supporting 58 percent of the center's budget while industry and vendors account for 22 percent of the costs. The Commonwealth of Pennsylvania and the National Institutes of Health both support PSC, accounting for 8 percent and 4 percent of budget respectively. Excluding working students, the center has a staff of around 65.

San Diego Supercomputer Center

The San Diego Supercomputer Center (SDSC) is located on the campus of the University of California at San Diego and is operated by General Atomics. SDSC is linked to 25 consortium members but has a user base in 44 States. At the end of 1988, over 2,700 users were accessing the center. SDSC has 48 industrial partners who use the facility's hardware, software, and support staff.

A Cray X-MP/48 was installed in December 1985. SDSC's first upgrade, a Y-MP8/864, was planned for

December 1989. In addition to the Cray, SDSC has five Sun workstations, two IRIS workstations, an Evans and Sutherland terminal, five Apollo workstations, a Pixar, an Ardent Titan, an SCS-40 minisupercomputer, a Supertek S-1 minisupercomputer, and two Symbolics machines.

The University of California at San Diego spends more than $250,000 a year on utilities and services for SDSC. For fiscal year 1990 the SDSC believes NSF will account for 47 percent of the center's operating budget. The State of California currently provides $1.25 million per year to the center and in 1988 approved funding of $6 million over 3 years to SDSC for research in scientific visualization. For fiscal year 1990 the State is projected to support 10 percent of the center's costs. Industrial support, which has given the center $12.6 million in donations and in-kind services, is projected to provide 15 percent of the total costs of SDSC in fiscal year 1990.

Other High-Performance Computer Facilities

Before 1984 only three universities operated supercomputers: Purdue University, the University of Minnesota, and Colorado State University. The NSF supercomputing initiative established five new supercomputer centers that were nationally accessible. States and universities began funding their own supercomputer centers, both in response to growing needs on campus and to increased feeling on the part of State leaders that supercomputer facilities could be important stimuli to local R&D and, therefore, to economic development. Now, many State and university centers offer access to high-performance computers (HPC);[1] and the NSF centers are only part of a much larger HPC environment including nearly 70 Federal installations (see table A-1).

Supercomputer center operators perceive their roles in different ways. Some want to be a proactive force in the research community, leading the way by helping develop new applications, training users, and so on. Others are content to follow in the path that the NSF National Centers create. These differences in goals/missions lead to varied services and computer systems. Some centers are "cycle shops," offering computing time but minimal support staff. Other centers maintain a large support staff and offer consulting, training sessions, and even assistance with software development. Four representative centers are described below.

Minnesota Supercomputer Center

The Minnesota Supercomputer Center, originally part of the University of Minnesota, is a for-profit computer center owned by the University of Minnesota. Currently, several thousand researchers use the center, over 700 of which are from the University of Minnesota. The Minnesota Supercomputing Institute, an academic unit of the university, channels university usage by providing grants to the students through a peer review process.

Table A-1—Federal Unclassified Supercomputer Installations

Laboratory	Number of machines
Department of Energy:	
Los Alamos National Lab	6
Livermore National Lab, NMFECC	4
Livermore National Lab	7
Sandia National Lab, Livermore	3
Sandia National Lab, Albuquerque	2
Oak Ridge National Lab	1
Idaho Falls National Engineering	1
Argonne National Lab	1
Knolls Atomic Power Lab	1
Bettis Atomic Power Lab	1
Savannah/DOE	1
Richland/DOE	1
Schenectedy Naval Reactors/DOE	2
Pittsburgh Naval Reactors/DOE	2
Department of Defense:	
Naval Research Lab	1
Naval Ship R&D Center	1
Fleet Numerical Oceanography	1
Naval Underwater System Command	1
Naval Weapons Center	1
Martin Marietta/NTB	1
Air Force Weapons Lab	2
Air Force Global Weather	1
Arnold Engineering and Development	1
Wright Patterson AFB	1
Aerospace Corp	1
Army Ballistic Research Lab	2
Army/Tacom	1
Army/Huntsville	1
Army/Kwajalein	1
Army/WES (on order)	1
Army/Warren	1
Defense Nuclear Agency	1
National Aeronautics and Space Administration:	
Ames	5
Goddard	2
Lewis	1
Langley	1
Marshal	1
Department of Commerce:	
National Institute of Standards and Technology	1
National Oceanic and Atmospheric Administration	4
Environmental Protection Agency:	
Raleigh, North Carolina	1
Department of Health and Human Services:	
National Institutes of Health	1
National Cancer Institute	1

SOURCE: Office of Technology Assessment estimate, September 1989.

[1] The number cannot be estimated exactly. First, it depends on the definition of supercomputer one uses. Secondly, the number keeps changing as States announce new plans for centers and as large research universities purchase their own HPCs.

The Minnesota Supercomputer Center received its first machine, a Cray 1A, in September 1981. In mid-1985, it installed a Cyber 205; and in the latter part of that year, two Cray 2 computers were installed within 3 months of each other. Minnesota bought its third Cray 2, the only one in use now, at the end of 1988, just after it installed a ETA-10. The ETA-10 has recently been decommissioned due to the closure of ETA. A Cray X-MP has been added, giving the center a total of two supercomputers. The Minnesota Supercomputer Center has acquired more supercomputers than anyone outside the Federal Government.

The Minnesota State Legislature provides funds to the university for the purchasing of supercomputer time. Although the university buys a substantial portion of supercomputing time, the center has many industrial clients whose identities are proprietary, but they include representatives of the auto, aerospace, petroleum, and electronic industries. They are charged a fee for the use of the facility.

The Ohio Supercomputer Center

The Ohio Supercomputer Center (OSC) originated from a coalition of scientists in the State. The center, located on Ohio State University's campus, is connected to 20 other Ohio universities via the Ohio Academic Research Network (OARNET). As of January 1989, three private firms were using the center's resources.

In August 1987, OSC installed a Cray X-MP/24, which was upgraded to a Cray X-MP/28 a year later. In August 1989 the center replaced the X-MP with a Cray Research Y-MP. In addition to Cray hardware, there are 40 Sun Graphic workstations, a Pixar II, a Stallar Graphics machine, a Silicon Graphic workstation, and an Abekas Still Store machine. The center maintains a staff of about 35..

The Ohio General Assembly began funding the center in the summer of 1987, appropriating $7.5 million. In March 1988, the Assembly allocated $22 million for the acquisition of a Cray Y-MP. Ohio State University has pledged $8.2 million to augment the center's budget. As of February 1989 the State has spent $37.7 million in funding.[2] OSC's annual budget is around $6 million (not including the purchase/leasing of their Cray).

Center for High Performance Computing (CHPC)

The Center for High Performance Computing is located at the University of Texas at Austin. CHPC serves all 14 institutions, 8 academic institutions, and 6 health related organizations, in the University of Texas system.

The University of Texas installed a Cray X-MP/24 in March 1986, and a Cray 14se in November 1988. The X-MP is used primarily for research. For now, the Cray 14se is being used as a vehicle for the conversion of users to the Unix system. About 40 people staff the center.

Original funding for the center and the Cray X-MP came from bonds and endowments from both the University of Texas system and the University of Texas at Austin. The annual budget of CHPC is about $3 million. About 95 percent of the center's operating budget comes from State funding and endowments. Five percent of the costs are recovered from selling CPU time.

Alabama Supercomputer Network

The George C. Wallace Supercomputer Center, located in Huntsville, Alabama, serves the needs of researchers throughout Alabama. Through the Alabama Supercomputer Network, 13 Alabama institutions, university, and government sites are connected to the center. Under contract to the State, Boeing Computer Services provides the support staff and technical skills to operate the center. Support staff are located at each of the nodes to help facilitate the use of the supercomputer from remote sites.

A Cray X-MP/24 arrived in 1987 and became operational in early 1988. In 1987 the State of Alabama agreed to finance the center. The State allocated $2.2 million for the center and $38 million to Boeing Services for the initial 5 years. The average yearly budget is $7 million. The center has a support staff of about 25.

Alabama universities are guaranteed 60 percent of the available time at no cost, while commercial researchers are charged a user fee. The impetus for the State to create a supercomputer center has been stated as the technical superiority a supercomputer would bring, which would draw high-tech industry to the State, enhance interaction between industry and the universities, and promote research and the associated educational programs within the university.

Commercial Labs

A few corporations, such as the Boeing Computer Corp., have been selling high performance computer time for a while. Boeing operates a Cray X-MP/24. Other commercial sellers of high performance computing time include the Houston Area Research Center (HARC). HARC operates the only Japanese supercomputer in america, the NEC SX2. The center offers remote services.

Computer Sciences Corp. (CSC), located in Falls Church, Virginia, has a 16-processor FLEX/32 from Flexible Computer Corp., a Convex 120 from Convex Computer Corp., and a DAP210 from Active Memory Technology. Federal agencies constitute two-thirds of

ne Ware, "Ohioans: Blazing Computer," *Ohio*, February 1989, p.12.

CSC's customers.[3] Power Computing Co., located in Dallas, Texas, offers time on a Cray X-MP/24. Situated in Houston, Texas, Supercomputing Technology sells time on its Cray X-MP/28. Opticom Corp., of San Jose California, offers time on a Cray X-MP/24, Cray 1-M, Convex C220, and C1 XP.

[3] Norris Parker Smith, "More Than Just Buying Cycles," *Supercomputer Review*, April 1989.

Index

U. S. GOVERNMENT PRINTING OFFICE ; 1991 - 33-982 QL 3

Superintendent of Documents **Publications** Order Form

Order Processing Code
***6993**

P3

Charge your order.
It's easy!

MasterCard VISA

To fax your orders and inquiries—(202) 275-0019

□ **YES,** please send me the following indicated publications:

________ copies of ***Seeking Solutions: High-Performance Computing for Science (20 pages),*** S/N 052-003-01227-8 at $2.25 each.

________ copies of ***Summary (20 pages),*** S/N 052-003-01226-0 at $1.00 each.

□ Please send me your **Free Catalog** of hundreds of bestselling Government books.

The total cost of my order is $________. (International customers please add 25%.) Prices include regular domestic postage and handling and are good through 10/91. After this date, please call Order and Information Desk at 202-783-3238 to verify prices.

____
(Company or personal name) (Please type or print)

____
(Additional address/attention line)

____
(Street address)

____
(City, State, ZIP Code)

( ) ____________________
(Daytime phone including area code)

Please Choose Method of Payment:

□ Check payable to the Superintendent of Documents

□ GPO Deposit Account □□□□□□□-□

□ VISA or MasterCard Account

□□□□□□□□□□□□□□□□□□□□

____
(Credit card expiration date)

Thank you for your order!

____
(Signature) 4/91

Mail To: Superintendent of Documents, Government Printing Office, Washington, DC 20402-9325