Reports Leading to the National High Performance Computing Program

"A Research and Development Strategy for High Performance Computing", November 1987

"The U.S. Supercomputer Industry,"
December 1987*

"The Federal High Performance Computing Program",
September, 1989.

"High Performance Computing and Communication:
Investment in American Competitiveness", ("Gartner Report"),
March 1991.

A Research and Development Strategy for High Performance Computing

Executive Office of the President
Office of Science and Technology Policy
November 20, 1987

APPENDIX C: *HPC Strategy* REPRINT

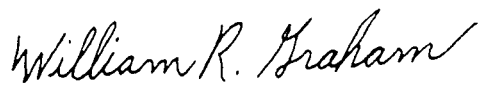
EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20506

This year the Federal Coordinating Council for Science, Engineering, and Technology (FCCSET) Committee on Computer Research and Applications began a systematic review of the status and directions of high performance computing and its relationship to federal research and development. The Committee held a series of workshops involving hundreds of computer scientists and technologists from academia, industry, and government. A result of this effort is the report that follows, containing findings and recommendations concerning this critical issue. It has been sent to the appropriate committees of Congress for their review.

A consistent theme in this report is the need for industry, academia, and government to collaborate and exchange information on future R&D efforts. Partners need to give one another signals as to their intent for future activities, and this report is a necessary first step in that process. The vision it represents must continue to grow. For that reason, I have asked the Committee to initiate the appropriate forums for discussing it further with the computing community.

Another theme has come out of this report: within four decades, the field of computer science has moved from a service discipline to a pervasive technology with a rigorous scientific basis. Computer science has become important to our national security and to our industrial productivity, and as such it provides the United States with many opportunities and challenges. Three of those opportunities are addressed in the report's findings and recommendations: High Performance Computers, Software Technology and Algorithms, and Networking. The fourth recommendation involves the Basic Research and Human Resources that will be required to conduct the other initiatives.

One thing is clear: the competition in an increasingly competitive global market cannot be ignored. The portion of our balance of trade supported by our high performance computing capability is becoming more important to the nation. In short, the United States must continue to have a strong, competitive supercomputing capability if it is to remain at the forefront of advanced technology. For that reason the Office of Science and Technology Policy is encouraging activities among the federal agencies together with the academic community and the private sector.



William R. Graham
Science Adviser to the President and
Director, Office of Science and Technology Policy

APPENDIX C: *HPC Strategy* REPRINT

CONTENTS

SUMMARY OF FINDINGS ON COMPUTER RESEARCH AND APPLICATIONS

1. HIGH PERFORMANCE COMPUTERS: **A strong domestic high performance computer industry is essential for maintaining U.S. leadership in critical national security areas and in broad sectors of the civilian economy.**

   - o U.S. high performance computer industry leadership is challenged by government supported research and development in Japan and Europe.
   - o U.S. leadership in developing new component technology and applying large scale parallel architectures are key ingredients for maintaining high performance computing leadership. The first generation of scalable parallel systems is now commercially available from U.S. vendors. Application-specific integrated circuits have become less expensive and more readily available and are beginning to be integrated into high performance computers.

2. SOFTWARE TECHNOLOGY AND ALGORITHMS: **Research progress and technology transfer in software and applications must keep pace with advances in computing architecture and microelectronics.**

   - o Progress in software and algorithms is required to more fully exploit the opportunity offered by parallel systems.
   - o Computational methods have emerged as indispensable and enabling tools for a diverse spectrum of science, engineering, design, and research applications.
   - o Interdisciplinary research is required to develop and maintain a base of applications software that exploits advances in high performance computing and algorithm design in order to address the "grand challenges" of science and engineering.

3. NETWORKING: **The U.S. faces serious challenges in networking technology which could become a barrier to the advance and use of computing technology in science and engineering.**

   - o Current network technology does not adequately support scientific collaboration or access to unique scientific resources. At this time, U.S. commercial and government sponsored networks are not coordinated, do not have sufficient capacity, do not interoperate effectively, and do not ensure privacy.
   - o Europe and Japan are aggressively moving ahead of the U.S. in a variety of networking areas with the support of concentrated government and industry research and implementation programs.

4. BASIC RESEARCH AND HUMAN RESOURCES: **Federal research and development funding has established laboratories in universities, industry, and government which have become the major sources of innovation in the development and use of computing technology.**

SUMMARY OF RECOMMENDATIONS FOR A NATIONAL HIGH PERFORMANCE COMPUTING STRATEGY

1. HIGH PERFORMANCE COMPUTERS: The U.S. Government should establish a long range strategy for Federal support for basic research on high performance computer technology and the appropriate transfer of research and technology to U.S. industry.

2. SOFTWARE TECHNOLOGY AND ALGORITHMS: The U.S. should take the lead in encouraging joint research with government, industry, and university participation to improve basic tools, languages, algorithms, and associated theory for the scientific "grand challenges" with widespread applicability.

3. NETWORKING: U.S. government, industry, and universities should coordinate research and development for a research network to provide a distributed computing capability that links the government, industry, and higher education communities.

4. BASIC RESEARCH AND HUMAN RESOURCES: Long term support for basic research in computer science should be increased within available resources. Industry, universities, and government should work together to improve the training and utilization of personnel to expand the base of research and development in computational science and technology.

APPENDIX C: *HPC Strategy* REPRINT

A Research and Development Strategy for High Performance Computing

High performance computing refers to the full range of supercomputing activities including existing supercomputer systems, special purpose and experimental systems, and the new generation of large scale parallel architectures.

The Challenge

In the span of four decades, computing has become one of the most pervasive and powerful technologies for information management, communications, design, manufacturing, and scientific progress.

The U.S. currently leads the world in the development and use of high performance computing for national security, industrial productivity, and science and engineering, but that lead is being challenged. Through an increased foreign industrial capability, the U.S. technology lead in computing has diminished considerably in recent years, but the U.S. continues to maintain strength in basic science and technology. The technology is changing rapidly and the downstream rewards for leadership are great. Progress in computing can be accelerated through the continued pioneering of new hardware, software, algorithms, and network technology and the effective transition of that technology to the marketplace. A shared computing research and development vision is needed to provide to government, industry, and academia a basis for cooperative action. The successful implementation of a strategy to attain this vision and a balanced plan for transition from one generation of technology to the next can result in continued strength and leadership in the forthcoming decades.

High performance computing technology has also become essential to progress in science and engineering. A **grand challenge** is a fundamental problem in science or engineering, with broad applications, whose solution would be enabled by the application of the high performance computing resources that could become available in the near future. Examples of grand challenges are: (1) Computational fluid dynamics for the design of hypersonic aircraft, efficient automobile bodies, and

extremely quiet submarines, for weather forecasting for short and long term effects, efficient recovery of oil, and for many other applications; (2) Electronic structure calculations for the design of new materials such as chemical catalysts, immunological agents, and superconductors; (3) Plasma dynamics for fusion energy technology and for safe and efficient military technology; (4) Calculations to understand the fundamental nature of matter, including quantum chromodynamics and condensed matter theory; (5) Symbolic computations including speech recognition, computer vision, natural language understanding, automated reasoning, and tools for design, manufacturing, and simulation of complex systems. Many of these could be considerably advanced by the use of computer systems capable of trillions of operations per second.

THE STRATEGY

A **High Performance Computing Strategy**, involving close coordination of existing programs and augmented effort, is required to address this national challenge. This strategy involves the coordinated pursuit of computing technology goals through joint efforts of government, industry, and academia. The strategy will have impact in clarifying and focusing the direction of Federally-funded computing research, which continues to be the major source of innovation for computing technology and a primary catalyst for industrial development. Government support should be highly leveraged with resources provided by industry participants. To be effective, the strategy should also be defined and continually updated in cooperation with industry and academia by making them participants in developing and implementing a shared vision of the future to ensure continued U.S. leadership.

The high performance computing strategy is designed to sustain and focus basic Federally-funded research and promote the transfer of basic science from the laboratory to U.S. industrial development and finally to the marketplace. Technology development will be encouraged as appropriate to meet immediate needs as well as to create a foundation for long term leadership. Strong emphasis will be placed on continued transfer of the results of government funded R&D to industry and on cooperation with industry to insure the continued strength of American high technology trade in the international marketplace.

The basic elements of the strategy are research and development programs in high performance computer architecture, in custom hardware, in software and algorithms, and in networking technology, all supported by a basic research foundation. In each of these areas, major opportunities exist that require coordinated support and management, building on existing government programs. Access to high performance computing is essential for providing scientists and engineers at research institutions throughout the country with the ability to use the most advanced computers for their work. The strategy needs to concurrently address the appropriate Federal role in each

of the basic elements of the R&D process—basic research, applied research, and industrial development—in order to meet long term, intermediate, and short term technology development goals. Explicit attention must be directed to the flow of technology from basic to applied areas and to the marketplace, as well as back into the research community to create the next generation of computing infrastructure, achieving a cumulative effect. Technology developments within individual element areas will contribute extensively to other activities. Simultaneous and coordinated pursuit of the areas is therefore an important element of the strategy.

CURRENT STATUS AND TRENDS

• **High performance computing systems.** Improvements in materials and component technology are rapidly advancing computer capability. Memory and logic circuits are continuing to improve in speed and density, but as fundamental physical limits are approached, advances are being sought through improved computer architectures, custom hardware, and software. Computer architecture has begun to evolve into large scale multiple processor systems, and in the past four years a first generation of scalable parallel systems has progressed from the research laboratory to the marketplace. Scalable architectures provide a uniform approach that enables a wide range of capacity, from workstations to very high performance computers. Application-specific integrated circuits, such as for real-time signal processing, are being incorporated into special purpose computers.

At current performance levels our ability to model many important science, engineering, and economic problems is still limited. Formulations of computational models presently exist that for realistic solutions would require speeds of teraflops (trillions of floating point operations per second) and equivalent improvement in memory size, mass storage, and input/output systems. In addition, symbolic processing is complementing and enhancing numeric approaches. Achievement of this performance level in the next 5 years appears to be a feasible goal, based on credible extrapolations of processor capability, number of processors, and software sophistication. In developing the new architectural approaches, however, careful collaboration will be required with the applications community to assess the various approaches and to achieve transition to the new approaches where appropriate. As transitions are made, the high performance computing industry should strive to maintain its continued leadership and competitveness.

• **Software technology and algorithms.** As high performance computing systems evolve and become more critical in science, engineering, and other applications domains, software technology becomes an increasingly central concern. As experienced in many U.S. space and defense programs, for example, software can become the dominant computational cost element in large systems because of the need to support evolution throughout the system life cycle from design and

development to long term maintenance and transition to the next generation. Future software environments and tools should support the development of trustworthy systems capable of evolution, while increasing productivity of developers and users of the systems. Effective exploitation of the performance potential of the emerging parallel systems poses a special challenge both to software and to algorithm design.

High performance computing offers scientists and engineers the opportunity to use computer models to simulate conditions difficult or impossible to create and measure in the laboratory. This new paradigm of computational science and engineering offers an important complement to traditional theoretical and experimental approaches, and it is already having major impact in many areas. New approaches combining numeric and symbolic methods are emerging. The development of new instruments and data generation methods in fields as diverse as genetics, seismology, and materials accelerates demand for computational power. In addition, the opportunity is created to coordinate and focus effort on important grand challenges, such as computational fluid dynamics, weather forecasting, plasma dynamics, and other areas.

• **Computer network technology.** A modern high speed research network is one of the elements needed to provide high performance distributed computation and communication support for research and technology development in government, academia, and industry. A coordinated research network based on very high bandwidth links would enable the creation of large-scale geographically distributed heterogeneous systems that link multiple high performance workstations, databases, data generation sources, and extremely high performance servers as required, in order to provide rapid and responsive service to scientists and engineers distributed across the country. The existing national network is a collection of loosely coupled networks, called an internet, based on concepts pioneered in the U.S.

Technical issues being addressed include utilization of fiber optics to improve performance for the entire research and higher education enterprise of the nation. An additional issue of pressing concern, particularly within the governmental and industrial sectors, is that of computer and network security to ensure privacy and trustworthiness in a heterogeneous network environment. At present, responsibility for privacy and the assurance of trust are vested principally in the computers and switching nodes on the network. Further research, already actively underway, is urgently needed to develop models, methodology, algorithms and software appropriate to the scale of a coordinated research network.

• **Basic research and human resources in Computer and Computational Science.** Federal funding has historically been, and will likely remain, a major source of support for important new ideas in computing technology. Carefully managed and stable funding is required to maintain vigorous research in computer and computational science and sufficient growth in computer science manpower. It is important to maintain the strength of the existing major research centers and to develop new research activity to support the growth in computer and computational science. Interactions should be fostered among academia, industry, and national laboratories to address large problems and to promote transfer of technology. In the longer term, enhancement of the computing technology base will have significant impact in productivity, efficiency, and effectiveness of government, industry, and the research community.

IMPACT

Computing technology is vital to national security. Advanced computer systems and software are now integral components in most major defense, intelligence, and aerospace systems. Computing technology has a central role in energy research, oil exploration, weapons research, aircraft design, and other national security technology areas.

Major advances in science and engineering have also accrued from recent improvements in supercomputing capability. The existence of machines with hundred megaflop (hundreds of millions of floating point operations per second) speed and multimillion word memories has allowed, for the first time, accurate treatment of important problems in weather prediction, hydrodynamics, plasma physics, stress analysis, atomic and molecular structure, and other areas. The emerging machines with 1 to 10 gigaflop (billions of flops) speed and 100 to 300 million word memories are expected to produce comparable advances in solving numeric and symbolic problems.

Many of these advances in science and engineering are the result of the application of high performance computing to execute computational simulations based on mathematical models. This approach to science and engineering is becoming an important addition to traditional experimental and theoretical approaches. In applications such as the National Aerospace Plane, supercomputing provides the best means to analyze and develop strategies to overcome technical obstacles that determine whether the hypersonic vehicle can fly beyond speeds of Mach seven, where wind tunnels reach their maximum capability. The list of applications for which supercomputing plays this kind of role is extensive, and includes nearly all high-technology industries. The extent of its usage makes supercomputing an important element in maintaining national competitiveness in many high technology industries.

The high performance computing strategy will have impact in many sectors of the economy. Nearly all sectors of advanced industry are dependent on computing infrastructure. Any improvement in computing capability will have substantial leveraged impact in broad sectors, particularly as applications software increases in power and sophistication.

The computer hardware industry alone amounted to $65 billion in 1986, and U.S. technical market dominance, long taken for granted, is now challenged in this and other critical areas, including networking, microsystems and custom high-performance integrated circuit technology. Foreign investment in computing research and technology has grown considerably in the last decade.

As stated in the report of the White House Science Council, *Research in Very High Performance Computing*, November 1985, "The bottom line is that any country which seeks to control its future must effectively exploit high performance computing. A country which aspires to military leadership must dominate, if not control, high performance computing. A country seeking economic strength in the information age must lead in the development and application of high performance computing in industry and research."

BACKGROUND

The Federal Coordinating Council on Science, Engineering and Technology (FCCSET) was established by Congress under the Office of Science and Technology Policy (OSTP) to catalyze interagency consideration of broad national issues and to coordinate various programs of the Federal government. The FCCSET in turn, established a series of committees, with interagency participation to assess and recommend action for national science and technology issues. The committees have become recognized as focal points for interagency coordination activity, addressing issues that have been identified by direct requests through the OSTP and indirect requests by member agencies (such as the NSF requirement to provide an update to the Lax Report on Large Scale Computing in Science and Engineering). These studies have enabled the FCCSET Committee on Computer Research and Applications to develop a national view of computing technology needs, opportunities, and trends.

From its inception, the FCCSET Committee on Supercomputing (the original name of this committee) was chartered to examine the status of high performance computing in the U.S. and to recommend what role the Federal Government should play regarding this technology. The committee issued two reports in 1983 that provided an integrated assessment of the status of the supercomputer industry and recommended government actions. The FCCSET Committee on Computer Research and Applications concluded that it would be proper to include an update of the earlier reports to address the changes that have occurred in the intervening period as a complement to the technical

reports. The review was based upon periodic meetings with and site visits to supercomputer manufacturers and consultation with experts in high performance scientific computing. White papers were contributed to this report by industry leaders and supercomputer experts. The report was completed in September 1987 and its findings and recommendations are incorporated in the body of this report.

In developing the recommendations presented in this report, the FCCSET Committee reviewed findings and recommendations from a variety of sources, including those mentioned above. A related activity has been the preparation by the White House Science Council (WHSC) Committee on Research in Very High Performance Computing of the report *Research in Very High Performance Computing*, November 1985. The WHSC Committee, composed of respected individuals from academia, industry, and government, made recommendations related to the issues more recently addressed by the FCCSET Committee. In the areas addressed by both committees, there is a significant consistency of recommendations, and, indeed, progress in recent months further strengthens the case for the recommendations. The convergence of views expressed in the many reports, the strong interest in many sectors of government in developing a policy, the dramatic increase in foreign investment and competitiveness in computing and network technology, and the considerable progress in computing technology development worldwide are all indicators of the urgency of developing and implementing a strategy for nationwide coordination of high performance computing under the auspices of the government.

One of the of the direct requests that this report responds to is in Public Law 99-383, August 21, 1986, in which Congress charged the Office of Science and Technology Policy to conduct a study of critical problems and of current and future options regarding communications networks for research computers, including supercomputers, at universities and federal research facilities in the United States. The legislation asked that requirements for supercomputers be addressed within one year and requirements for all research computers be addressed within two years. Dr. William R. Graham, Director of the Office of Science and Technology Policy, subsequently charged the Federal Coordinating Council on Science Engineering and Technology (FCCSET) Committee on Computer Research and Applications to carry out the technical aspects of the study for OSTP.

It was recognized by the FCCSET Committee on Computer Research and Applications that networking technology needs to be addressed in the context of the applications of computing and the sources of computing power that are interconnected using the network technology. This report, therefore, presents an integrated set of findings and recommendations related to Federal support for computer and related research.

Three subcommittees carried out the work. Each of these committees contributed to the Findings and Recommendations contained in this report. The result is an integrated set of recommendations that addresses the technical areas.

- The **Subcommittee on Computer Networking, Infrastructure, and Digital Communications** invited experts in government, industry and academia to write white papers on networking trends, requirements, concepts applications, and plans. A workshop involving nearly 100 researchers, network users, network suppliers, and policy officials was held in San Diego, California in February 1987 to discuss the white papers and to develop the foundation for the report. Workshop leaders and other experts later met in Washington to summarize the workshop discussions and focused on six topics: access requirements and future alternatives, special requirements for supercomputer networks, internet concepts, future standards and services requirements, security issues, and the government role in networking. As a result of this work, the participants recommended that no distinction should be made between networks for supercomputers and other research computers and that the final report to the Congress should address networks generally. The requirements for both supercomputers and for other research computers are, therefore, addressed in this report.

- The **Subcommittee on Science and Engineering Computing** assessed computing needs related to computational science and engineering. The committee focused its deliberations on requirements for high performance computing, on networking and access issues, and on software technology and algorithms. Under the auspices of the Society for Industrial and Applied Mathematics (SIAM), and with the support of NSF and DOE, a workshop involving 38 recognized leaders from industry, academia, and national laboratories was held at Leesburg, Virginia in February 1987 on research issues in large-scale computational science and engineering. This workshop focused on advanced systems, parallel computing and applications. As a result of the workshop report, recommendations were made related to the role of computing technology in science and engineering applications.

- The **Subcommittee on Computer Research and Development** assessed the role of basic research, the development of high performance computing technology, and issues related to software technology. Contributing to this activity were two workshops. The National Science Foundation (NSF) Advisory Committee for Computer Research reviewed the field and produced an Initiatives Report in May 1987. This report recommended investment in three areas, including parallel systems and software technology. In September 1987, the Defense Advanced Research Projects Agency (DARPA) held a workshop on advanced computing technology in Gaithersburg, Maryland involving 200 researchers from academia, industry, and government. The workshop focused on large-scale parallel systems and software approaches to achieving high performance computing.

An important result of the activity of the FCCSET Committee on Computer Research and Applications and its subcommittees is that increased coordination among the Government elements is necessary to implement a strategy for high performance computing. The findings and recommendations presented here represent a consensus reached among the subcommittees and convey the powerful and compelling vision that emerged. As a result of this process, the next step would be for the members of the Committee on Computer Research and Applications to develop a plan to help ensure that the vision is shared between government, academia, and American industry. Subsequently, the Committee should develop an implementation plan for Federal government activities, including a detailed discussion of overall priorities.

1. High Performance Computers

- **FINDING: A strong domestic high performance computer industry is essential for maintaining U.S. leadership in critical national security areas and in broad sectors of the civilian economy.**

U.S. prominence in technology critical to national defense and industrial competitiveness has been based on leadership in developing and exploiting high performance computers. This preeminence could be challenged by dependency upon other countries for state of the art computers. Supercomputer capability has contributed for many years to military superiority. In addition, industrial applications now constitute more than half of the supercomputer market and are an important factor in U.S. industrial competitiveness. However, continued progress in computational science and engineering will depend in large part on the development of computers with 100 to 1000 times current capability for important defense, scientific, and industrial applications. These applications are represented by the grand challenges.

- **U.S. high performance computer industry leadership is challenged by government supported research and development in Japan and Europe.**

The U.S. currently leads the world in research, development, and use of supercomputers. However, this leadership faces a formidable challenge from abroad, primarily from the Japanese. The 1983 FCCSET report stated that "The Japanese have begun a major effort to become the world leader in supercomputer technology, marketing, and applications." Most of the analyses and projections advanced in support of that statement have proven to be accurate.

Japanese supercomputers have entered the marketplace with better performance than expected. Japanese supercomputer manufacturers have attained a high level of excellence in high speed, high density logic and memory microcircuits required for advanced supercomputers. As a result, some U.S. computer manufacturers are dependent on their Japanese competitors for sole supply of critical microcircuits. Japanese manufacturers, universities, and government have demonstrated the ability to cooperate in developing and marketing supercomputers as well as in advancing high performance computing. Recent successes in dominating related high-technology markets underscore their financial, technical, and marketing capability.

- **U.S. leadership in developing new component technology and applying large scale parallel architectures are key ingredients for maintaining high performance computing leadership. The first generation of scalable parallel systems is now commercially available from U.S. vendors. Application-specific integrated circuits have become less expensive and more readily available and are beginning to be integrated into high performance computers.**

  The current generation of supercomputers achieve their performance through the use of the fastest possible individual components, but with relatively conservative computer architectures. While these computers currently employ up to eight parallel processors, their specific architectures cannot be scaled up significantly. Large scale parallel processing, in which the computational workload is shared among many processors, is considered to be the most promising approach to producing significantly faster supercomputers. The U.S. is currently the leader in developing new technology as well as components. However, exploiting these techniques effectively presents significant challenges. Major effort will be required to develop parallel processing hardware, algorithms, and software to the point where it can be applied successfully to a broad spectrum of scientific and engineering problems.

  Government funded R&D in universities and industry has focused on an approach to large-scale parallelism that is based on aggressive computer architecture designs and on high levels of circuit integration, albeit with somewhat slower individual components. Unlike current supercomputers, the resulting systems employ 100s to 10,000s of processors. Equally important, these architectures are scalable to higher levels of parallelism with corresponding increase in potential performance.

  The first generation of scalable parallel systems is now commercially available from U.S. vendors. These systems have demonstrated high performance for both numeric and non-numeric, including symbolic processing. Comparable systems do not yet exist outside the U.S. The second generation, with higher speed individual components and more parallelism, is already in development here. Experience with these systems has shown that, even with existing software, they are effective for certain classes of problems. New approaches to software for these large-scale parallel systems are in the process of emerging. These approaches suggest that parallel architecture may be effective for wide classes of scientific and engineering problems. An important benefit of the scalable architectures is that a single design, with its attendant components and software, may prove to be useful and efficient over a performance range of 10 to 100 or more. This allows one design to be used for a family of workstations, mini-supercomputers, and supercomputers.

- **RECOMMENDATION: The U.S. Government should establish a long range strategy for Federal support for basic research on high performance computer technology and the appropriate transfer of research and technology to U.S. industry.**

The program should build upon existing government supported efforts. However, government funding should not be viewed as a substitute for private capital in the high performance computer marketplace. A primary objective is to ensure continued availability of domestic sources for high performance computers that are required for Federal programs, both civilian and defense. These actions should include:

- Government should support, when appropriate for mission requirements, the acquisition of prototype or early production models of new high performance computers that offer potential for improving research productivity in mission areas. These computers could be placed in centers of expertise in order to allow sophisticated users to share initial experiences with manufacturers and other users, and to develop software to complement the vendor's initial offerings. These initial acquisitions should not require the vendor to supply mature operating systems and applications software typical of production computers. However, a criterion for acquisition should be that the hardware designs reflect a sensitivity to software issues, and that the computer has the potential for sustained performance in practical applications that approaches the peak hardware performance.

- Government agencies should seek opportunities to cooperate with industry in jointly funded R&D projects, concentrating especially on those technologies that appear scalable to performance levels of trillions of operations per second (teraops) for complex science, engineering, and other problems of national importance. Systems are needed for both numeric and symbolic computations.

However, since government mission requirements typically exceed those of industrial applications, cooperating with industry in R&D for computers to meet these missions will help to assure that the necessary computers are available. It will also drive supercomputer development at a faster pace than would be sustained by commercial forces alone, an important factor retaining and increasing U.S. leadership in this area.

- Government agencies should fund basic research to lay the foundation for future generations of high performance computers. Steps should be taken to ensure that development of state-of-the-art computers continues to be monitored for appropriate export controls.

2. Software Technology and Algorithms

- **FINDING: Research progress and technology transfer in software and applications must keep pace with advances in computing architecture and microelectronics.**

  - **Progress in software and algorithms is required to more fully exploit the opportunity offered by parallel systems.**
  - **Computational methods have emerged as indispensable and enabling tools for a diverse spectrum of science, engineering, and design research and applications.**
  - **Interdisciplinary research is required to develop and maintain a base of applications software that exploits advances in high performance computing and algorithm design in order to address the "grand challenges" of science and engineering.**

A **grand challenge** is a fundamental problem in science and engineering, with broad application, whose solution will be enabled by the application of the high performance computing resources that could become available in the near future.

As high performance computing systems evolve and are applied to more challenging problems, it is becoming increasingly clear that advances in software technology and applications are essential to realize the full performance potential of these systems. Software development, analysis, and adaptation remain difficult and costly for traditional sequential systems. Large scale complex systems including parallel systems pose even greater challenges. Market pressures for the early release of new computing system products have created a tradition of weak systems software and inadequate programming tools for new computers.

Current approaches to software development provide only limited capabilities for flexible, adaptable, and reusable systems that are capable of sustained and graceful growth. Most existing software is developed to satisfy nearer term needs for performance at the expense of these longer term needs. This is particularly the case for applications in which specific architectural features of computers have been used to obtain maximum performance through low level programming techniques. The lack of portability of these programs significantly raises the cost of transition to newer architectural approaches in many applications areas. Approaches are beginning to emerge in the research community that have a potential to address the reuse and portability problems.

Experiments with parallel computers have demonstrated that computation speeds can increase almost in direct proportion to the number of processors in certain applications. Although it is not yet possible to determine in general the most

efficient distribution of tasks among processors, important progress has nonetheless been made in the development of computational models and parallel algorithms for many key problem areas.

Access to advanced computing systems is an important element in addressing this problem. Experience has shown that the quality of systems and applications software increases rapidly as computing systems are made more available. Initial generic operating systems and extensions to existing programming languages can provide access through coupling high performance computers with existing workstations using either direct or network connections. However, in order to achieve the full potential impact of large scale parallel computing on applications, major new conceptual developments in algorithms and software are required.

The U.S. leads in many areas of software development. The Japanese, however, also recognize the need for high quality software capability and support in order to develop and market advanced machines. They have demonstrated the ability to effectively compete, for example in the area of sophisticated vectorizing compilers.

The U.S. will need to encourage the collaboration of computer scientists, mathematicians, and the scientists in critical areas of computing applications in order to bring to bear the proper mix of expertise on the software systems problem. Such collaboration will be enhanced by network technology, which will enable geographically dispersed groups of researchers to effectively collaborate on "grand challenges." Critical computer applications include problems in fluid dynamics, plasma physics, elucidation of atomic and molecular structure, weather prediction, engineering design and manufacturing, computer vision, speech understanding, automated reasoning, and a variety of national security problems.

- **RECOMMENDATION: The U.S. should take the lead in encouraging joint research with government, industry, and university participation to improve basic tools, languages, algorithms, and associated theory for the scientific "grand challenges" with widespread applicability.**

  Software research should be initiated with specific focus on key scientific areas and on technology issues with widespread applicability. This research is intended to accelerate software and algorithm development for advanced architectures by increased early user access to prototype machines. It would also provide settings for developing advanced applications for production machines. Software technology needs to be developed in real problem contexts to facilitate the development of large complex and distributed systems and to enable transition of emerging parallel systems technology into the computing research community and into the scientific and engineering applications communities.

  As part of a mixed strategy, longer term and more basic software problems of reliability and trust, adaptability, and programmer productivity must continue to be addressed. Languages and standards must be promoted that permit development of systems that are portable without sacrificing performance.

  In applications areas including computational science and engineering, technology should be developed to support a smooth transition from the current software practice to new approaches based on more powerful languages, optimizing compilers, and tools supported by algorithm libraries. The potential of combining symbolic and numeric approaches should be explored. Progress in these areas will have significant impact on addressing the "grand challenges" in computational science and engineering. Although there are many pressing near term needs in software technology, direct investment in approaches with longer term impact must be sustained if there is to be significant progress on the major challenges for software technology while achieving adequate system performance.

  Applications include (1) distributed access to very large databases of scientific, engineering, and other data, (2) high bandwidth access to and linking among shared computational resources, (3) high bandwidth access to shared data generation resources, (4) high bandwidth access to shared data analysis resources, such as workstations supporting advanced visualization techniques.

3. NETWORKING

- **FINDING: The U.S. faces serious challenges in networking technology which could become a barrier to the advance and use of computing technology in science and engineering.**

  - **Current network technology does not adequately support scientific collaboration or access to unique scientific resources. U.S. commercial and government sponsored networks presently are not coordinated, do not have sufficient capacity, do not interoperate effectively, and do not ensure privacy.**
  - **Europe and Japan are aggressively moving ahead of the U.S. in a variety of networking areas with the support of concentrated government and industry research and implementation programs.**

Computer network technology provides the means to develop large scale distributed approaches to the collaborative solution of computational problems in science, engineering, and other applications areas. Today, researchers sharing a local area network are able to exploit nearly instantaneous communication and sharing of data, creating an effect of linking their workstations and high performance servers into a single large scale heterogeneous computing facility. This kind of capability is now appearing in larger scale campus-wide computer networks, enabling new forms of collaboration. National networks, on the other hand, have low capacity, are overloaded, and fail to interoperate successfully. These have been expanded to increase the number of users and connections but the performance of the underlying network technology has not kept pace with the increased demands. Therefore, the networks which in the 1970s had significant impact in enabling collaboration, are now barriers. Only the simplest capabilities, such as electronic mail and small file transfers, are now usable. Capacity, for example, is orders of magnitude less than the rates required, even if the network is used only for graphics.

Other countries have recognized the value of national computing networks, and, following the early U.S. lead, have developed and installed national networks using current technology. As a result, these countries are now much better prepared to exploit the new opportunities provided by distributed collaborative computing than the U.S. is at the present time. The basic technologies for later generations are also being developed in the U.S., but there have been no major efforts to apply them to address the needs.

A longer term goal is the creation of large scale geographically distributed heterogeneous systems that link multiple superworkstations and high performance supercomputers to provide service to scientists and engineers distributed across the country. A well-coordinated national network could link these resources together when required on an *ad hoc* basis to provide rapid response to computational needs as they arise. This could reduce the number of sites needed for the physical presence of supercomputers. Present access to computer networks by researchers is dependent upon individual funding or location. There is unnecessary duplication in the links from various agencies to each campus. The development of improved networking facilities could greatly stimulate U.S. research and provide equitable access to resources.

Many scientific research facilities in the U.S. consist of a single, large, and costly installation such as a synchrotron light source, a supercomputer, a wind tunnel, a particle accelerator, or a unique database. These facilities provide the experimental apparatus for groups of scientific collaborators located throughout the country. Wide area networks are the logical mechanism for making data from such facilities more easily accessible nationwide. An important issue is that of computer and network security to ensure privacy and trustworthiness in a heterogeneous network environment. At present, responsibility for privacy and the assurance of trust are vested principally in the computers and switching nodes on the network.

Existing government-supported wide-area networks include ARPANET, HEPNET, MFENET, NSFNET, NASNET, MILNET, and SPAN, as well as private and commercial facilities such as TYMNET, TELENET, BITNET, and lines leased from the communication carriers. Longer-range estimates vary, but it is expected that by the year 1995 the nation's research community will be able to make effective use of a high capacity national network with capacity measured in billions of bits per second. Without improved networks, speed of data transmission will be a limiting factor in the ability of researchers to carry out complex analyses. The digital circuits most widely available today with transmission speeds of 56 kilobits per second (kb/s) are impediments to leading edge research and to optimal remote high performance computer use.

Point-to-point connections require interconnects through multiple vendors with cumulative costs. Greater network speed can reduce the time required to perform a given experiment and increase both the volume of data and the amount of detail that can be seen by researchers. Scientists accessing supercomputers would benefit because access speed is often critical in their work. Improved functionality frees scientists to concentrate directly on their experimental results rather than on operational details of the network. Increased network size extends these opportunities to thousands of individuals at smaller academic institutions throughout the nation. These modernization measures would significantly enhance the nation's position in scientific research. A national network would help maintain the U.S. leadership position in computer architectures,

microprocessors, data management, software engineering, and innovative networking facilities, and promote the development of international networking standards based on U.S. technology.

Integrated Systems Digital Networks (ISDN–voice and data) have been installed abroad on a national or regional scale. Research abroad is being conducted on service up to 1 Gb/s. Within the next five years, Integrated Services Digital Network (ISDN) circuits ranging from 64 kb/s to 1.5 Mb/s will be available in the larger metropolitan areas of the U.S. However, these services will fall short of the requirements for computer networks. By 1988 more than fifty Campus Area Networks will be operational at speeds approaching 100 Mb/s. Wide area networks operating at 1.5 Mb/s or less will not be able to handle the data volume expected.

Japan and Europe have extensive efforts with experimental nets in intermediate (40Mb) and high (gigabit) range. Japan is studying operational aspects of fiber nets using their national research network as a testbed, which includes exploring the feasibility of fiber optic services to residences.

To estimate the network bandwidth needed to support research at a major installation, the kinds and volume of traffic that would be used have been estimated at a representative campus, extrapolated ten years into the future. Three models were used to compute three independent estimates of the requirements for bandwidth needed by type of work, information needs by type of user, and information flowing at the installation boundary. In each model, the peak bandwidth was estimated for each type of service. For example, in the Task model, the need is dominated by that of at least one researcher to receive full color and full-motion high resolution images. A high-resolution color image contains about 30 megabits of information, so that a display rate of 30 frames per second requires a bandwidth of nearly one gigabit per second (Gb/s). In the User model, a research university with 35,000 students and 3,000 faculty and research staff using a mix of bandwidths again requires an aggregate bandwidth of approximately one Gb/s. In the Edge of the Installation model, bandwidth is estimated by the types of remote facilities being accessed and the expected number of simultaneous users; typical facilities include particle accelerators, supercomputers, and centers for imaging and/or animation. The aggregate bandwidth needed is one Gb/s. Thus three independent means of estimating bandwidth arrive at nearly the same requirement for a large research installation, and one Gb/s can confidently be used as a lower bound on the bandwidth of a national research network.

- **RECOMMENDATION:** **U.S. government, industry, and universities should coordinate research and development for a research network to provide a distributed computing capability that links the government, industry, and higher education communities.**

A research network should be established in a staged approach that supports the upgrade of current facilities and development of needed new capabilities. Achievement of this goal would foster and enhance the U.S. position of world leadership in computer networking as well as provide infrastructure for collaborative research. The FCCSET Committee on Computer Research and Applications should provide a forum for interagency cooperations. Elements of the plan should include:

- *Stage 1.* Upgrade existing facilities in support of a transition plan to the new network through a cooperative effort among major government users. The current interagency collaboration in expanding the Internet system originated by DARPA should be accelerated so that the networks supported by the agencies are interconnected over the next two years.
- *Stage 2.* The nation's existing networks that support scientific research should be upgraded and expanded to achieve data communications at 1.5 Mb/sec for 200 to 300 U.S. research institutions.
- *Stage 3.* Develop a system architecture for a national research network to support distributed collaborative computation through a strong program of research and development. A long-term program is needed to advance the technology of computer networking in order to achieve data communication and switching capabilities to support transmission of three billion bits per second (3 Gb/s) with deployment within fifteen years.
- Develop policy for long term support and upgrading of current high performance facilities, including timetables for backbone and connection development, industry participation, access, agency funding, tariff schedules, network management and administration. Support should be given to the development of standards and their harmonization in the international arena.

Until the national research network can replace the current system, existing networks should be maintained and modified as they join the national network. Remedial action should be initiated as soon as possible. Upgrading the backbone to at least 1.5 Mb/s should be accomplished by 1990. This will ensure that the new generation of high performance computing can be effectively interconnected.

Industry should be encouraged to participate in research, development, and deployment of the national research network. Telecommunication tariff schedules

which have been set for voice transmission should be reviewed in light of the requirements for transmission of data through computer networking.

Prompt effective coordination is needed to increase user participation in the standards development process, to get requirements for standards expressed early in the development process, and to speed the implementation of standards in commercial off-the-shelf products. It is essential that standards development be carried out within the framework of overall systems requirements to achieve interoperability, common user interfaces to systems, and enhanced security.

4. Basic Research and Human Resources

- **FINDING: Federal research and development funding has established laboratories in universities, industry, and government which have become the major sources of innovation in the development and use of computing technology.**

  Many of the advances in computer science and technology in the U.S. were made possible by Federal programs of research support to universities and industry. For example, the advances that have occurred since 1983 in the area of parallel computing are the direct result of Federal research investment through agencies including DARPA and NSF. In the area of application of supercomputers to science and engineering, the majority of this investment came from the NSF Advanced Scientific Computing centers. In the area of parallel architectures, the major investment came from the DARPA Strategic Computing Program. Programs sponsored by DOE, NASA, and Defense to support critical mission needs have been a major source of investment in computational applications research. In industry, support for basic research is only a small fraction of industry research most of which is focused on nearer term product development. This can be attributed in part to the long term and high risk nature of basic research, but a more significant inhibitor of investment is the difficulty in the computer industry of maintaining proprietary protection for certain kinds of key fundamental advances.

- **RECOMMENDATION: Long term support for basic research in computer science should be increased within available resources. Industry, universities, and government should work together to improve the training and utilization of personnel to expand the base of research and development in computational science and technology.**

  Maintain vigorous research in Computer Science and sufficient growth of computer science manpower to support the scientific/technological basis of the computer field. Foster interactions among academia, industry, and national laboratories by creating interdisciplinary teams to address large scale problems. Extend the technology base to attain significant impact on competitiveness and industrial productivity.

  Innovative very high performance computing systems should be made available to universities and basic research laboratories in order to assist in the evaluation and exploitation of new technology and new industrial innovations.

Continue the following successful approaches to basic research and development: (1) The practice of loosely coordinated and flexible basic research supported through various federal sectors and applied to a diversity of institutions, (2) The mixed strategy of peer review to support a broad range of exploratory basic research throughout the academic community and the complementary technical program management approach of larger scale experimental systems programs which exploit new opportunities as they emerge, (3) Support for individuals and small groups in theoretical areas, (4) The practice of supporting the relevant basic research as part of larger experimental systems projects.

IMPLEMENTATION

Success of the National High Performance Computing Strategy will require an attitude of cooperation in which academia, industry and government work effectively together in developing and assessing new technology and in achieving the transition of promising new ideas into the marketplace. The rapid pace of developments in computing technology creates a number of implementation challenges that must be addressed explicitly if the Strategy is to have maximum impact.

The FCCSET Committee on Computer Research and Applications provides an appropriate forum for coordination of Federal agency programs. Specifically:

- The subcommittee on Computer Networking, Infrastructure, and Digital Communications will develop a coordinated implementation plan for the national research network.

- The subcommittee on Science and Engineering Computing will review the *grand challenges* through the use of high performance computing systems, including the research that will be involved.

- The subcommittee on Computer Research and Development will review the need for advanced software, algorithms, and hardware for future high performance computing systems.

All of the subcommittees will consider appropriate action to secure a foundation of basic research and human resources. In all three subcommittees we expect some overlap of responsibility and interchange of ideas to be compatible with success.

As has been firmly stated, the full cooperation through a shared vision between government, industry and the research community will be a necessary ingredient for the successful implementation of this strategy. The FCCSET Committee on Computer Research and Applications therefore calls for timely consideration of the vision and strategy by representative bodies of the research community and industry.

It is essential, however, that implementation of the strategy be undertaken in a timely manner. There is a need to follow through on the breakthroughs that occurred partially as a result of federal investment in the early 1980s. The fast pace of development dictates that appropriate Federal efforts are needed to help ensure continued excellence in high speed networking technology and leadership in high performance computing. Foreign investment in technology development in these key areas has increased dramatically. The prudent strategy is to maintain a consistent strong lead in research and to transfer the results as quickly as possible to American industry.

COST ESTIMATES

Many of the basic elements of the high performance computing strategy are already being implemented as part of ongoing agency programs at DOE, DARPA, NSF, NASA, and other Federal agencies, and important progress is being made. The FCCSET Committee activity has contributed to achieving a shared vision, and early coordination is already occurring in anticipation of implementation of the strategy. Implementation of the strategy involves three principal funding components, including the national research network, joint research to address the "grand challenges," and basic research in high performance computing architecture, custom hardware design, software, algorithms, and supporting technologies. Multiple agencies are involved in the implementation and funding of each of these components.

The funds that would be associated with each of these components are described below. Obviously, any incremental funding must be evaluated and approved within the context of current activities and research needs in other high priority fields. Currently, the Federal government is spending about $500M per year on all aspects of high

Summary of Additional Funds
(Millions of Dollars)

Current Funds			Yr 1	Yr 2	Yr 3	Yr 4	Yr 5
50 [a]	National Research Network	Stage 1	5	5	5	0	0
		Stage 2	5	5	55	55	55
		Stage 3	40	40	40	40	40
150 [b]	Joint Research in Computational Science and Engineering		30	60	90	120	150
300	Basic Research in Computer Science and High Performance Computing		60	120	180	240	300
500	TOTAL (above current funds)		140	230	370	455	545
	Funding Increase by Year (noncumulative)		140	90	140	85	90

a Estimated network research and support in grants and contracts.
b Estimated operating costs for existing computational science facilities.

performance computing. Funding for the activities recommended in this report would increase this base by $140M in additional resources for the first year, growing to an additional $545M per year in 5 years.

National Research Network. Current operating costs for the present collection of research-support networks operated by DARPA, NSF, DOE, and NASA is approximately $50M per year; the figure is uncertain because many subnetworks are funded by increments on research grants and contracts, rather than being centrally supported. Currently the interconnection of existing agencies' networks is planned within existing budgets. A significant increase in investment is needed to achieve the required capability. This investment could occur in three concurrent stages.

The *first stage* activity would involve an immediate upgrade to 1.5 Mbit/sec of the existing research-support networks. This would cost $15M over three years.

The *second stage* would expand upgraded network services (45Mbit/sec) to 200 to 300 research installations, using primarily fiber-optic trunk facilities. Development costs for this stage would be $5M per year of additional funding. Operation of the upgraded network would commence in three to five years, with operating costs of approximately $50M per year. Since the transition from the first stage to the second stage network could not be instantaneous, initially the full operating cost of the second stage network would necessitate additional funding; that requirement will diminish to the extent that the first stage network is phased out.

The goal of the *third stage* would be to deliver one to three Gbit/sec to selected research facilities, and 45 Mbit/sec to approximately 1000 research sites. Research and development costs for this project are estimated at $400M of new funds, spent over ten years; after ten years, operating costs would be about $200M per year unless some tariff relief is achieved.

Joint Research in Computational Science and Engineering. Current operating costs for existing computational science laboratory facilities is approximately $150M per year. Additional investment would be required to upgrade the existing facilities and/or to establish additional joint research activities, with government, industry, and university participation, to address approximately specific problem areas, including selected *grand challenges*. Many of these joint research efforts will involve multiple physical sites connected by the research network. The investment in these research activities supports pursuit of the grand challenges. This includes personnel to develop computational approaches in terms of theory, algorithms, and software, and the acquisition of modern computing equipment. Estimated Federal costs average $15M per year to establish and sustain each grand challenge. The joint research activities would be introduced at the rate of two per year. Overall investment will be approximately $30M per year initially, increasing to $150M per year in five years as new grand challenges are added.

Basic Research in Computer Science and High Performance Computing. Current Federal investment in advanced computer research is estimated at $300M in FY88. Over the past four years, investment in these areas has grown at 15% per year. The rate of increase appears to be declining, however, at a time when increased investment appears to be needed. Sufficient resources should continue to be allocated to take full advantage of the high performance computing opportunities that now exist including design and prototype development of systems capable of trillions of operations per second. A second important element is stable funding, which is required to preserve the long-term strength of the research community.

Other countries are also devoting considerable resources in this area. For example, the Japanese government supports two projects which directly address supercomputer development: The Fifth Generation Project and the Superspeed Project. Support for each of these is estimated to be in excess of $100M per year. In addition to this government support, Japanese industry is investing considerably more to develop high performance computers. Japanese government and industry are also investing amounts comparable to those recommended here to develop high bandwidth research networks.

ACKNOWLEDGMENTS

Office of Science and Technology Policy guidance was provided by Michael Marks. Stephen L. Squires, Defense Advanced Research Projects Agency, acted as Executive Secretary for this report. Technical assistance was provided by William L. Scherlis, Defense Advanced Research Projects Agency; along with Kathleen Bernard, Office of Science and Technology Policy; Charles N. Brownstein, National Science Foundation; Leslie Chow, National Aeronautics and Space Administration; and Michael Crisp, Department of Energy.

FCCSET COMMITTEE ON COMPUTER RESEARCH AND APPLICATIONS

Paul G. Huray (Chair)
Office of Science and Technology Policy

SUBCOMMITTEES

Science and Engineering Computing

James F. Decker (Chair)
Department of Energy

James Burrows
National Bureau of Standards

John S. Cavallini
Health and Human Services

Melvyn Ciment
National Science Foundation

John Connolly
National Science Foundation

Craig Fields
Defense Advanced Research Projects Agency

Harlow Freitag
Supercomputer Research Center

Randolph Graves
National Aeronautics and Space Administration

Norman H. Kreisman
Department of Energy

Lewis Lipkin
National Institutes of Health

Allan T. Mense
Strategic Defense Initiative Office

David B. Nelson
Department of Energy

C. E. Oliver
Air Force Weapons Lab

John P. Riganati
Supercomputer Research Center

Paul B. Schneck
Supercomputer Research Center

K. Speierman
National Security Agency

Computer Research and Development

Saul Amarel (Chair)
Defense Advanced Research Projects Agency

Donald Austin
Department of Energy

C. Gordon Bell
National Science Foundation

James Burrows
National Bureau of Standards

Bernard Chern
National Science Foundation

Peter Freeman
National Science Foundation

Lee Holcomb
National Aeronautics and Space Administration

Charles Holland
Office of Naval Research

Robert E. Kahn
Computer Science Technology Board

Daniel R. Masys
National Institutes of Health

Robert Polvado
Central Intelligence Agency

David Sadoff
Department of State

William L. Scherlis
Defense Advanced Research Projects Agency

K. Speierman
National Security Agency

Stephen L. Squires
Defense Advanced Research Projects Agency

Charles F. Stebbins
Air Force Systems Command

Daniel F. Weiner, II
Joint Tactical Fusion Program

Computer Networking, Infrastructure and Digital Communications

C. Gordon Bell (Chair)
National Science Foundation

Ronald Bailey
National Aeronautics and Space Administration

Sandra Bates
National Aeronautics and Space Administration

James Burrows
National Bureau of Standards

John S. Cavallini
Health and Human Services

Thomas Kitchens
Department of Energy

James Oberthaler
National Institutes of Health

Dennis G. Perry
Defense Advanced Research Projectes Agency

Arnold Pratt
National Institutes of Health

Shirley Radack
National Bureau of Standards

Rudi F. Saenger
Naval Research Laboratory

Daniel VanBelleghem
National Science Foundation

Stephen Wolff
National Science Foundation

DOE/ER-0362

Federal Coordinating Council on Science, Engineering and Technology

Committee on Computer Research and Applications
Subcommittee on Science and Engineering Computing

The U.S. Supercomputer Industry

December 1987

Office of Science and Technology Policy
Executive Office of the President
Washington, DC 20506

FCCSET Report
on
The U.S. Supercomputer Industry

Executive Summary

i. Introduction

The Federal Coordinating Council on Science, Engineering, and Technology (FCCSET) Committee on Supercomputing was chartered by the Director of the Office of Science and Technology Policy in 1982 to examine the status of supercomputing in the United States and to recommend a role for the Federal Government in the development of this technology. This FCCSET Committee issued two reports[2,3] in 1983 and one[7] in 1985 that recommended government actions necessary for the continued development and use of supercomputers in the United States. An important input to the committee's deliberations was the Report of the Panel on Large Scale Computing in Science and Engineering, or Lax Report,[1] that provided an integrated assessment of supercomputing in these disciplines.

In the study that follows, the FCCSET Committee (now called the Subcommittee on Science and Engineering Computing of the FCCSET Committee on Computer Research and Applications) reports on the status of the supercomputer industry and addresses changes that have occurred since issuance of the 1983 and 1985 reports. The review has been based upon periodic meetings with and site visits to supercomputer manufacturers and consultation with experts in high performance scientific computing. White papers have been contributed to this report by industry leaders and supercomputer experts.

ii. FCCSET found that:

- **A vigorous domestic supercomputer industry is essential for maintaining U.S. leadership in critical defense areas and in areas important for our civilian economy**

U.S. preeminence in many critical technology areas has been based on leadership in developing and exploiting supercomputers. This leadership would be jeopardized by dependency upon other countries for state-of-the-art supercomputers. Government use of supercomputers has spawned industrial uses that confer competitive advantage to the user. Industrial applications continue to constitute more than half of the supercomputer market and are an important factor in U.S. industrial competitiveness. This is partly a consequence of effective technology transfer from the Federal laboratories that use supercomputers.

- **U.S. supercomputer leadership is threatened**

The U.S. currently leads the world in research, development and use of supercomputers. However, this leadership faces a formidable challenge from abroad, primarily from the Japanese. U.S. supercomputer manufacturers are small when compared with the giant Japanese vertically integrated companies that have targeted supercomputers as one of their future growth areas. Japanese markets are difficult for U.S. supercomputer manufacturers to penetrate. Furthermore, the U.S. manufacturers are perilously dependent on their Japanese competitors for critical semiconductor chips.

- **The Federal Government has retreated from its historic role as "friendly buyer" of supercomputers**

There has been a decline in the Federal role of "friendly buyer" in which Government agencies would acquire the first prototypes of innovative new computers, even if there were no software available. Initial use of these prototypes demonstrated the viability of new supercomputers and helped to establish their acceptability in the marketplace. Now, manufacturers of new supercomputers find that at times they must develop complete suites of software, even for initial sales to government laboratories. This has slowed the introduction of new supercomputers and has lengthened the time required for manufacturers to see a return on their investment.

- **A strong federal role in ensuring US leadership in supercomputers is justified by national security needs and the needs of federal research programs**

Nearly half of all domestic supercomputers are still used for federal programs. National Security programs must not grow dependent on foreign sources for the fastest supercomputers. A healthy domestic supercomputer industry will foster US international competitiveness.

- **U.S. leadership in parallel processing is a key ingredient for maintaining supercomputer leadership**

Parallel processing architecture is the most promising approach to producing significantly faster supercomputers. The U.S. is currently the leader in

the development of parallel processing hardware and software. Exploiting parallelism effectively presents formidable challenges. If properly employed, U.S. skill and experience in these areas can help to preserve national supercomputing leadership.

- **Impacts on the supercomputer industry are not thoroughly considered in formulating trade policy**

Some aspects of U.S. trade policy have had unintended negative impacts on the U.S. supercomputer industry. The recent U.S./Japan semiconductor chip agreement has resulted in increases in some prices of chips for which the supercomputer industry has no other source.

National security controls placed upon the export of supercomputers and related products, as well as delays in issuing export licenses, have placed U.S. companies at a competitive disadvantage with respect to foreign manufacturers in certain non-Eastern Bloc markets.

iii. FCCSET Recommends that:

1. The U.S. Government carry out a coordinated long-range R&D program in supercomputer applications, software, and advanced computer architectures

— A primary objective is to ensure continued development of supercomputers by the private sector that are required for use in Federal programs. Current national defense and basic science and engineering programs could benefit significantly from faster supercomputers. The program should also be based on an anticipation of future needs.

— The program should build upon existing government-supported efforts.

— Major scientific and engineering challenges should serve as a focus for the effort.

— Advanced computer architectures should be developed.

— A 1000 fold improvement in applied computational capability in five years should be a short-term goal.

— R&D should address algorithms, both systems and applications software, and peripherals for future supercomputers.

— The program will have the beneficial by-product of aiding the development of supercomputers for industrial applications.

2. The Federal Government should return to its role as a "friendly" buyer of innovative supercomputers.

Government agencies should budget for acquisition of prototype or "serial one" models of new supercomputers that offer potential for improving their research productivity. These initial acquisitions should not require complete operating systems and applications software typical of production computers.

3. The Federal Government should make federally developed software available to the private sector when possible

Manufacturers should be expected to develop software for production computers, and software developed by government laboratories for prototype computers should be available to the private sector within the constraints of national security requirements.

I. Background and Historical Perspective

Starting with the effort to build the Enigma computer that would solve encryption problems during World War II, most of the early motivation for high performance computing has centered on the need to solve large mathematical, scientific and engineering problems for national security. In 1943 the Electronic Numerical Integrator and Calculator (ENIAC) was built by Mauchly and Eckert at the University of Pennsylvania to solve ballistics problems for the Army. In the 1950's and 1960's national security requirements stimulated supercomputer developments by IBM, Burroughs, Texas Instruments, and Control Data Corporation. During the 1960's and 1970's, serial number one of most supercomputers was placed in the Department of Energy's national laboratories for the design of nuclear weapons. The government played a significant role

in encouraging the development of new supercomputers by actively encouraging and supporting supercomputer manufacturers.

Today the problems that require supercomputers have grown in importance and number (see References 13 and 14 for selected applications). There is a direct relationship between U.S. leadership in supercomputers and a strong national defense. Supercomputers play a vital role in designing nuclear and directed energy weapons, in designing aerospace vehicles, and in handling many problems related to command, control, communications, and intelligence. From calculating artillery trajectories in 1943 to modern battle management/target acquisition problems, supercomputers have contributed to the defense mission and today are essential. The technological edge the U.S. enjoys in its weapons and intelligence systems cannot be maintained without U.S. leadership in supercomputers.

As supercomputers have evolved in the U.S., providing faster execution times for floating point operations and larger memories for storing data, a symbiotic relationship has emerged between government, industry, and universities to develop the entire computer system and its environment. Supercomputer manufacturers were concerned simply with the goal of providing hardware with faster processors and larger direct memory. Traditionally government, and more recently universities, worked to make the supercomputers cost effective to solve "real" problems. This implies developing models, algorithms, languages, compilers, and operating systems. Industry provided the necessary input/output devices to accommodate the new supercomputers.

Although the government has been the primary motivator and major buyer of supercomputers in the past, industry and university needs have recently grown to a roughly equal share of the market. The rise of this private sector market has provided additional incentives to computer vendors to build more powerful machines.

The aerospace and oil companies were the first in the industrial sector to acquire and use substantial numbers of the current class of supercomputers. The list of other industries has expanded to include: electronics, automobile, computer, chemical, and the motion picture industries. A growing number of companies now require supercomputers to be competitive.

Practically every area of science is finding the supercomputer to be an important tool in fundamental research.[13,14] In 1983, the National Science Foundation supported the establishment of five university supercomputer centers to provide supercomputer access to the basic research community. There are now 15 universities in the U.S. with high performance scientific computers providing much needed computing resources to the research community. As a result of the growth of industry and university users, the traditional U.S. supercomputer manufacturers (Cray and CDC/ETA) face the new challenge of providing a more complete system environment to customers who have little or no desire to develop software packages or operating systems. Thus, supercomputer manufacturers are finding it necessary to provide more software. This requires more resources than American manufacturers have spent in the past or may be capable of spending. Fortunately, U.S. manufacturers have been able to rely on software developed for the most part in the national laboratories, since the basic machine architectures have not changed dramatically in the past decade. However, manufacturers are planning more radical changes in the basic architecture (moving to much larger numbers of parallel processors) in order to achieve greater speeds and capability. The software and applications must be reformulated to use the full potential of the new hardware.

Because supercomputers are essential for U.S. national security, industrial competitiveness, and leadership in scientific research, a strong, self-sufficient, domestic supercomputer industry is vital to assure their continued development and availability.

From the first days of computing, the U.S. has been the leader in all aspects of high performance computing, research, development, and applications. This leadership is now threatened. With the 1983 arrival of NEC, Fujitsu, and Hitachi, three large, vertically integrated Japanese corporations, into the supercomputer arena, the attention to "systems" issues by the small U.S. manufacturers becomes more compelling.

II. Review of Previous Reports

Widespread concern developed during the early 1980's that the United States could lose its position of preeminence in the design, manufacture, and utilization of the largest scientific computers (supercomputers). This concern was expressed in a number of reports[1-7] and congressional hearings.[9,10]

A. Lax Report

The most widely distributed and referenced effort was the Report of the Panel on Large Scale Computing in Science and Engineering drawn from a June 1982 workshop led by Peter D. Lax of New York University.[1] The Lax Report recommended the establishment of a National Program to stimulate research, exploratory development, and expanded use of advanced computer technology. The program consisted of four components:

1. Increased access for the scientific and engineering research community through high bandwidth networks to an adequate number of state-of-the-art supercomputing facilities and experimental computers;
2. Increased research in computational mathematics, software, and algorithms necessary to the effective and efficient use of supercomputer systems;
3. Training of personnel in scientific and engineering computing;
4. Research and development basic to the design and implementation of new supercomputer systems of substantially increased capability and capacity, beyond that likely to arise from commercial requirements alone.

Underlying all four recommendations is the establishment of a system of effective computer networks that joins government, industrial, and university scientists and engineers. The Lax Panel recommended that this program be coordinated within the Federal government by an interagency policy committee and that an interdisciplinary Large Scale Computing Advisory Panel be established to assist in its planning, implementation, and operation.

B. FCCSET Reports

In January 1983, the Federal Coordinating Council on Science, Engineering, and Technology (FCCSET) responded to reports of threatened U.S. supercomputer leadership by forming the Panel on Supercomputers to examine the Government's role in the development and use of high performance scientific computers. As a result of discussions of this panel, three interagency working groups were formed. The **Procurement Panel** was asked to report on possible Government actions that could ensure U.S. supercomputer leadership. The **Access Panel** was to explore options for making government funded supercomputers more broadly available. The **Research Panel** was to report on coordination among government agencies that fund research effecting the supercomputer technology base.

1. Procurement Panel

The Procurement Panel found that while the Federal government was, in 1983, the largest user of supercomputers, the private sector was rapidly increasing its use. Furthermore, government programs required machines of significantly greater capacity than was available at the time. Historically, the Federal government had nurtured supercomputer development in the 1950's and 1960's through favorable procurement policies, direct funding of R&D, and providing software development. However, the level of Federal support declined in the 1970's. Furthermore, while early supercomputer development was characterized by strong Government-Industry-University interactions, the universities were out of the mainstream development in the 1970's.

The Procurement Panel concluded that U.S. leadership was necessary for national defense needs and to maintain economic competitiveness. To preserve the U.S. lead, the panel recommended:

- **Set a national goal of developing a computer 200 times faster than Class-VI machine and encouraging the industry to reach that goal through more favorable procurement policies;**
- **The Government reemphasize its past role of "friendly buyer" of the newest machines;**
- **Accelerate purchases of new machines to help the industry;**
- **Increase support of research; and**
- **Coordinate the entire program of supercomputer initiatives by establishing a permanent interagency group.**

2. Access Panel

The FCCSET Access Panel found that many federally funded facilities are operating at or near full capacity or have committed available capacity to future programmatic growth. Furthermore facilities that were supporting weapons or nuclear research would have to be totally restructured for unclassified

users to share computing resources. The three university supercomputer centers, at the time, were underutilized. These universities had difficulty obtaining the resources to provide the necessary, broad, user friendly services. The Access Panel found that high quality service should be a prime consideration when establishing supercomputer services for remote users. Finally they found that there were insufficient opportunities for training students in the physical sciences and engineering in the use of supercomputers and for training computer science students in the operation of supercomputer systems.

The FCCSET Access Panel agreed with the Lax Report that it was in the national interest to expand access to supercomputers and proposed to accomplish this with emphasis on the use of networks which would provide quality service to remote users. It was recommended that universities be brought back into the supercomputer mainstream through improved access and student support.

The Access Panel specifically recommended expansion of user community of existing supercomputers, initiation of access by common carriers for communications by remote users, coordination among agencies of networks, the establishment of new supercomputer centers and associated networks as needed, experimentation with high bandwidth communications systems, and the establishment of a mechanism for trading computer time between agencies.

3. Research Working Panel

The Research Working Panel issued a report on advanced computing research in June of 1985.[7] The panel recommended a vigorous and effectively coordinated federal research program. They recommended research in a variety of architectures with special emphasis on parallel processing. They also urged increased efforts in training of researchers and technology transfer of federally sponsored research. They recommended the development of performance modeling and measurement techniques. Subsequent reports were supportive of the conclusion and recommendations of the Lax and the FCCSET Panels' reports:

C. IEEE Report

The report of the IEEE Scientific Supercomputer Committee, chaired by Sidney Fernbach,[5] described the growing use of supercomputers in industry to replace traditional "experimental" methods of engineering. The report urged the Federal government to make a commitment to maintain U.S. leadership in supercomputers. They recommended direct support of research and the establishment of several supercomputer centers for research, teaching, and applications development. The Fernbach report recommended support of all technologies needed for supercomputer system development. They proposed tax incentives and antitrust relief for the U.S. industry to remain competitive. The Fernbach report also recommended the designation of a "lead-agency" to coordinate federal activities that impact the supercomputer industry.

D. President's Foreign Intelligence Advisory Board

The President's Foreign Intelligence Advisory Board (PFIAB)[6] recommended more support for the microchip development that provides the technological basis for much of the supercomputer industry as well as expanded efforts to gather data on progress in other countries (especially Japan). The report also recommended expanded government procurement coupled with algorithm and software research of advanced supercomputers.

E. Bardon-Curtis Report

The National Science Foundation (NSF) Working Group on Computers for Research[4] was organized to provide specific recommendations to be followed within the NSF and to propose budget plans. This report emphasized the academic perspective and recommended consideration of proposals for 10 new supercomputer systems, over 3 years, as one approach to improved access to supercomputers. The report also urged development of networks linking universities and laboratories with supercomputers.

III. Government Response to Reports

One of the immediate responses to the recommendations of the Lax Report and other reports was the establishment of several interagency panels, within the framework of the Federal Coordinating Council on Science, Engineering and Technology (FCCSET),

to serve as a forum to discuss supercomputer activities and issues throughout the Federal government. In 1983, two of the FCCSET Committee's Panel Reports (procurement and access panels) produced specific recommendations to implement the general recommendations of the Lax Report.

In the following year the Chairman of the FCCSET Committee requested from Committee members, brief reports that summarized the steps their respective agencies were taking in response to the recommendations of the FCCSET Procurement and Access Panels. The responses are summarized as follows:

	Procurement Panel	Access Panel
Department of Energy		
	FY 1985 request increase of $6M in Applied Mathematical Sciences to support university design and prototyping of several new research computers.	Provided availability of 5% of 2 Lawrence Livermore National Laboratory (LLNL) Crays to Magnetic Fusion Energy Computer (MFEC) network for new users.
		Requested an additional supercomputer in FY 1985 for above.
	Leased two Denelecor HEP-ls at Los Alamos National Laboratory (LANL) and Argonne National Laboratory (ANL) and two Elxsi machines at Sandia National Laboratory (SNL) and New York University (NYU) for research and experimentation in parallel processing.	Florida State University Supercomputer Program incorporated into DOE access program.
	Established Scientific Computing Staff to support research in applied mathematics and computer sciences.	Planned expansion of MFE Computer Center to service additional 1000-1500 users.
	Ordered a Cray II.	Scientific computing staff to manage MFE network.
National Aeronautics and Space Administration (NASA)		
	Established Numerical Aerodynamic Simulation (NAS) program in FY 1984 to provide national access to prototype and "serial number 1" supercomputers.	Transferring older computers to universities.
	Ordered a Cray II.	Installed NASA-wide high-speed network with tail circuits to 20 industry and university sites, and internetworking to DOD and university networks in 1986.
	Continue policy to order next generation supercomputer.	

Department of Commerce		
	Class VI computer operational at National Bureau of Standards (NBS) fall of 1985, to service NBS,	Time on new Class VI made available to universities in cooperative university/NBS/NOAA research.
National Oceanographic and Atmospheric Administration (NOAA):		NOAA Geophysical Fluid Dynamics Laboratory (GFDL) computer available to Princeton Univ. community.
		Conducting research in wideband satellite transmission between U.S. (NBS) and Germany.
National Science Foundation (NSF)		
	Plan to start one to three new computer centers in FY 1985.	To provide over 5000 hours of computer time (from Univ. of Minnesota, Purdue, Boeing Computer Sciences) in next 12 months. National NSF network to connect planned centers.
	Intent to establish five to ten such centers.	
Department of Defense		
(Air Force)		
	Fast Algorithm Initiative to support research to develop numerical methods and algorithms for parallel processing architectures funded FY 1985.	Providing access for Air Force Office of Scientific Research supported research to Air Force Weapons Laboratory Cray I.
		Attempting to enhance access of university researchers to AF supercomputers.
(Navy)		
		Office of Naval Research (ONR) to provide Naval Research Laboratory (NRL) computer time allocation of one shift for ONR contractors.
		ONR to head panel to review DOD support for research requiring supercomputer access and to coordinate with other agencies considering mechanisms for university access to supercomputers.

National Security Agency

Establishing Supercomputing Research Center in Maryland.

Planning support of IBM/NYU cooperative research/prototype project.

IV. Present Status of U.S. Supercomputer Industry

A. Changes since previous study

In January 1983, all of the approximately 60 supercomputers installed worldwide were of U.S. manufacture. By 1987, Japan had manufactured 57 of the world's 244 supercomputers. The remaining 187 were of the U.S. manufacture. Thus, the Japanese supercomputer market share has grown from zero in 1982 to over 23 percent by 1987. Cray Research Inc., continues to maintain market leadership in supercomputers but Cray's market share is eroding rapidly. Cray is healthy, profitable, and in 1986 did business at the rate of $600 million annually. Cray is the leading U.S. supercomputer manufacturer. The three Japanese supercomputer manufacturing companies are Fujitsu Ltd., Nippon Electric Corporation (NEC), and Hitachi Corporation. These are large vertically integrated companies with individual annual sales ranging from $6 billion to $20 billion.

Two supercomputer companies, in addition to Cray, were active in the U.S. at the time of the Lax and FCCSET reports. Control Data manufactured and marketed the Cyber 205 and held a 26 percent market share. Control Data established a subsidiary, ETA Systems, Inc., to develop and market the ETA-10 supercomputer, a candidate for 10 gigaflop performance. Production of the Cyber 205 has ceased, and the first installation of the ETA-10 began in late 1986.

Denelcor, Inc., went out of business in 1985. This small company had developed a promising parallel architecture supercomputer design, but lacked the resources required to establish a presence in the supercomputer market. Its demise indicates that substantial financial resources are necessary to maintain a long term presence in the market.

IBM was not a factor in the supercomputer market in 1983, although the company had marketed supercomputers in the 1960's and 70's. However, in 1987 IBM agreed to invest in Scientific Computer Systems, a new venture led by Steve Chen, formerly a Cray vice president in charge of X-MP/Y-MP efforts. Chen intends to create and market a 48/64 processor supercomputer with advanced architecture and components significantly superior to the present state-of-the-art within a five years. The relative vulnerability of Cray and ETA to their larger, emerging Japanese competitors in the supercomputer market place is discussed further in this report. However, a sustained effort by IBM, a $50 billion company with world-class R&D resources and the potential for exploiting advanced technology in hand or in prospect, could significantly effect the competitive equation. IBM has made such strategic investments in the past - for example purchasing approximately 25% of Intel Corporation, a major U.S. integrated circuit and microprocessor manufacturer. IBM has recently introduced vector processing capabilities to the 3090 line of mainframe computers, giving these systems marginal Class VI status. IBM has contributed a paper to this report discussing its corporate strategy vis-a-vis the supercomputer market.

World supercomputer sales in 1986 totaled nearly $900 million. With expected growth rates of 30-40 percent per year, the forecast for global shipments in 1990 is $2 billion.[15]

Historically, the market size has exceeded forecasts. In 1983, the Federal Coordinating Council on Science, Engineering and Technology (FCCSET) Panel on Supercomputers recommended government stimulation of U.S. development to achieve computer systems by 1990 of 200 times the power of Class VI machines. With the Class VI benchmark of 100 megaflops, the goal is within reach: The Cray XMP-4, on the same rating scale is one gigaflops; the Cray 2 is on the same order; the ETA-10, when design performance is achieved, will perform at 10 gigaflops. The Cray III and YMP are expected on the market before 1990.

B. Near Term Projections of Technology and Architecture vs. Need

Technology Projections

Improvements in the technology of materials and components are advancing computer capability. Speed and density are increasing in both memory and logic circuits. Forecasts are for 64-256 megabit chips in the 1990's. Faster and more efficient signal transmission resulting from emerging opto-electronic technologies is likely. However, the fundamental limits of physics remain. Transmission speeds will never exceed the speed of light. Semiconductor packaging density improvements become more difficult as feature sizes approach smaller multiples of atomic dimensions. Advanced fabrication techniques such as x-ray lithography, are probably limited to a 0.1 micron design rule. Given these physical limits on hardware, future advances are sought in supercomputer architectures and software. Architecture has begun to evolve into multiprocessor supercomputer systems. The XMP-2 and XMP-4 have 2 and 4 processors respectively; the Cray 2 has 4 processors; and the ETA-10 has 8 processors. It is expected that 16 or more processors will appear in near-term new systems. Certain special purpose machines have many parallel processors. The Connection Machine, developed for DARPA, has 65,536 one bit processors. Another route to speeding up computing is the "wide instruction" architecture, which decodes and executes several operations within a single instruction at once to achieve a very fine-grain parallelism. Appendix B describes a number of parallel processor computers and indicates some of the advantages and disadvantages with this type of architecture.

Experiments with multiprocessor machines have demonstrated that calculating speeds can increase almost in direct proportion to the number of processors for some applications. However, it is not yet possible to determine the optimum configuration of processors for all applications. Coordinating complex multiple memory and processor interaction to attack a complicated problem in science or engineering is challenging. It appears that very sophisticated software will be required to implement the potential of multiprocessor machines. However, most software today does not exploit the full capabilities of present single processor vector machines. The U.S. leads in many areas of software development; however, the Japanese also recognize the need for software capability and support in order to develop and market advanced machines. They have already produced superior vector compilers, based in part on U.S. research. In some markets of the computer industry, future software sales are expected to exceed those of hardware. Such may never be the case for supercomputers but (commercial) customer purchasing decisions will clearly be influenced by the software that is available for specific industry applications.

Increasingly, researchers require graphics software and graphics displays to interpret the results of supercomputer calculations. In some cases the only way to comprehend complex multidimensional phenomena is via graphics rather than by scanning columns of numbers. For some types of problem it is also necessary to observe the change over time in the simulated system, thus capability for motion pictures (animated graphical output) is increasing important. Interesting output systems have been created by interfacing a graphical workstation to the supercomputer. The workstation functions as a graphics postprocessor for the output generated by the supercomputer and drives a color graphics display that can present several high-resolution frames per second. Despite the attractiveness of this type of system, the technology is still in its infancy and will require much software development and system interfacing before it will be generally useful. Fortunately, the existence of standards for communications and graphics eases the interfacing problems.

Successful implementation of graphics displays will dramatically increase the requirements for supercomputer capability. Today it can take several minutes to compute a single frame of a complex hydrodynamics calculation. As researchers learn the value of animated displays they will demand closer to real-time simulations in order to run several cases in close succession.

Need

Reference 14 makes clear the advances in science and engineering that have accrued from past improvements in supercomputing capability. The existence of machines with hundred-megaflop speed and multimegaword memories has allowed, for the first time, accurate treatment of important problems in weather prediction, hydrodynamics, plasma physics, stress analysis, atomic and molecular structure, etc. The emerging machines with 1-10 gigaflop

speed and 100-300 megaword memories will produce similar advances.

Even at this performance level our ability to model important science, engineering, and economic problems will be limited. Model formulations already exist that would require teraflop (1000 gigaflops or 10^{12} flops) speeds and equivalent improvement in memory size for their solution. Fortunately, achievement of this performance level in the next five years appears to be a feasible goal, based on credible extrapolations in processor capability, number of processors, and software advances.[19]

C. Emergence of Mini-Supercomputers

Mini-supercomputers are entering the marketplace from a growing number of start-up and established firms. These new high speed machines, some designed for specific classes of problems, some using the same instruction sets as supercomputers, compete cost effectively with supercomputers for certain applications. They achieve a cost advantage over supercomputers by adapting less expensive, slightly lower performance microcircuits that were originally developed for other purposes. The proliferation of mini-supercomputers will make less clear the boundary between supercomputers and other computers. At least one supercomputer manufacturer plans to market mini-supercomputers, similar to their higher priced offerings, to address the lower end of the supercomputer market.

The introduction of mini-supercomputers is likely to expand sales of supercomputers. By acquiring mini-supercomputers, new customers can gain experience with supercomputer applications at lower initial cost. As their experience and processing requirements grow, they will likely migrate to higher performance supercomputers. Mini-supercomputers employing the same instruction set as true supercomputers offer the additional advantage of running exactly the same operating system, compilers, and applications codes. They are likely to be used as high-end stand alone workstations or may be interfaced to remote supercomputers for program debugging or initial runs using coarser computational grids. Experience gained from mini-supercomputers with novel architectures (especially those with large-scale parallelism) may indicate directions for future supercomputer architectures.

The proliferation of mini-supercomputers will complicate the implementation of U.S. export controls. It will be increasingly difficult to argue that there is a qualitative difference between supercomputers and lower .performance mini-supercomputers. Yet the relatively simple technology employed by some mini-supercomputers may make it difficult to control their export.

D. Problems Facing U.S. Vendors

U.S. supercomputer manufacturers face four major problems:

- **Dependency upon their foreign competitors for system components that may increase if the U.S. semiconductor industry or disk-storage industry declines.**
- **Aggressive Japanese trade practices.**
- **Burdensome U.S. controls on supercomputer exports.**
- **Large and growing capital resource requirements for developing next generation supercomputer systems.**

Component Problems

Cray Research and ETA Systems are relatively small companies that rely on external sources for semiconductor components and peripheral devices. In recent years this dependency has shifted from U.S. sources to Japanese suppliers. Lack of domestic self-sufficiency is a threat to future technical system design and manufacture. This growing dependency on foreign sources becomes more troubling when the foreign supplier is also an emerging competitor in the supercomputer market. The Japanese have a history of delaying or withholding technology in which they are leaders from their American competitors. For example, an American supercomputer manufacturer dependent upon a Japanese source for memory chips receives chips inferior to those the manufacturer provides for its own internal systems.

Similar cases have occurred in chip testers and in fabrication equipment. This has been a serious problem for U.S. manufacturers, whose efforts to remain in the lead require the latest and best component and fabrication technology.

American supercomputer manufacturers are well aware of the danger and have taken steps to reduce

their dependence on Japanese chip suppliers where possible. Cray declined Fujitsu offers to develop parts for the YMP (a successor to the XMP) and is working with Fairchild instead.[16] Fairchild was almost bought by Fujitsu. Cray is dependent on Fujitsu for much of the memory in the Cray 2. (The importance of this dependency is underscored by the fact that about one-third of the manufacturing cost of the Cray 2 is in integrated circuits.) ETA has achieved partial independence from Japanese suppliers by obtaining MOS gate arrays from Honeywell. Honeywell's capability in gate arrays derives from that company's participation in the Department of Defense's very high speed Integrated Circuit (VHSIC) R&D program.

The smaller U.S. firms are also at a disadvantage in acquiring computer peripherals, compared with the large vertically integrated Japanese manufacturers. Again, U.S. firms must source externally. More importantly, future developments in rotating (disk) storage are needed to match the capabilities of present and future supercomputers. Advances in candidate technologies such as optical and magneto-optical disks are needed in order to provide the tertiary storage and data transfer requirements for the next generation of supercomputers. Supercomputers are rapidly exceeding the capability of conventional magnetic recording technology to provide adequate input/output data rates. U.S. efforts to exploit new developments in magneto-optical disk technology are promising. However, the resources required to develop these new technology and associated manufacturing capability are high, while the market associated with supercomputers is small. There is little incentive for independent companies to take the risk, and the development is too expensive for small, domestic supercomputer manufacturers to finance.

Japanese Competitive Practices

The Government of Japan is engaged in a number of practices in the supercomputer industry which hurt the U.S. industry. For example, access by U.S. supercomputer vendors to the Japanese home market has been severely restricted. While industry analysts rate U.S. supercomputers as superior to the Japanese competition, only six such units have been sold in Japan, and none to a government supported agency or university. Seven years after establishing a sales and support office in Japan, Cray Research has sold six out of sixty-four supercomputers operating in Japan, whereas Cray has a majority of the supercomputer market in every other country. Domestic companies such as NEC, Fujitsu, and Hitachi appear to have the inside track when government agencies solicit bids for supercomputer contracts. The U.S. Trade Representative (USTR) has recently negotiated a supercomputer market access agreement with the Japanese government, but the practical effect on market share remains to be determined.

Export Controls on U.S. Manufactured Supercomputers

Supercomputers clearly fall into the category of advanced technology with potential for military applications. Their export from the U.S. to Soviet Bloc nations is controlled by the Coordinating Committee on Multilateral Export Controls (COCOM). Their export to non-aligned nations is controlled to avoid the possibility of reshipment to unfriendly nations or usage by unfriendly nations through access to the importing country. Even export to allied nations is controlled to assure that satisfactory user access controls are in place.

Rigid application of export controls to aggregate high technology exports from the US has been estimated to cost the US economy $9.3 billion annually.[17] Most of the lost sales are picked up by other Western nations. Approval of export licenses often takes more than 100 days. Supercomputers are subject to as much scrutiny as any other dual use high technology product and obtaining an export license usually requires several months. Even a supercomputer export to a western nation that is normally thought of as a U.S. ally is subject to time consuming review.

U.S. supercomputer manufacturers state that they are vulnerable to losing sales because foreign vendors obtain export licenses from their governments much faster and more easily. The foreign vendor can assure faster delivery and less interference with the customer's operations. To date no example has been cited for the loss of a US supercomputer sale to a foreign buyer due to excessive delays and red tape in the export control process. However, there is clearly considerable room for improvement and the current situation is a marketing hardship for American firms.

NATO members, as well as France and Japan, agree that militarily significant advanced technology should

not be transferred to the Communist Bloc and have cooperated through COCOM to prohibit such supercomputer sales. However, there is no agreed-to policy concerning export to non-aligned countries. Do we sell supercomputers to non-aligned countries, with risks of subsequent resale or military diversion? Should one distinguish between supercomputers and relatively ordinary computer technology available from sources in other countries? The development of mini-supercomputers and mainframes with vector processing capability allows militarily significant calculations to be carried out on much smaller, cheaper computers. Should their sale be prohibited or access to them controlled? It has proven to be extremely difficult to establish a workable policy, agreed to by all exporting countries, that would control the transfer of critical capability to potential adversaries.

E. U.S./Japan Semiconductor Trade Agreement

In September 1986, the U.S. and Japan signed an agreement on semiconductor trade. The chip accord forbid Japanese chip manufacturers from selling DRAMS below fair-market values, as calculated by the U.S. Department of Commerce. The values are supposed to reflect each Japanese producer's cost of production plus an 8 percent profit margin.

The short term impact of the agreement has been to raise the price to U.S. supercomputer manufacturers of chips that are available solely from Japanese sources. This puts U.S. produced supercomputers at a cost disadvantage compared with Japanese produced machines. As memory size grows and memory becomes a larger fraction of the total machine cost, these cost differences assume increasing importance in the competitive picture.

V. Analysis of Japanese Supercomputer Industry

A. Japanese Progress Since 1983

The 1983 FCCSET report stated that "The Japanese have begun a major effort to become the world leader in supercomputer technology, marketing, and applications." Most of the analyses and projects advanced in support of that statement have proven to be accurate. Japanese supercomputers have entered the marketplace with better performance than expected. Japanese supercomputer manufacturers have attained world leadership in high speed, high density logic and memory microcircuits required for advanced supercomputers. Japanese manufacturers, universities, and government have demonstrated the ability to cooperate in developing and marketing supercomputers. Because of their size, Japanese supercomputer manufacturers have much greater financial strength than their American counterparts and have already demonstrated their willingness to engage in "anticipatory pricing" (selling below cost) to gain entry into new markets.

In 1983 Fujitsu, Hitachi, and Nippon Electric Company (NEC) announced supercomputers with peak performance stated to exceed that of a Cray 1. The Fujitsu and Hitachi computers would have the added advantage to some customers of executing the IBM instruction set for compatibility with IBM systems and applications codes. By 1986 each of these computers had been installed and benchmarked, in some cases by American researchers running "real" applications codes as opposed to small demonstration kernels. Although the results differ from case to case because of varying hardware and compilers, in general the Japanese supercomputers are very competitive with their American counterparts.

The first supercomputers from each of the Japanese manufacturers were installed in national universities supported by Monbusho, the Ministry of Education, Science and Culture. The Fujitsu computer was installed at the Institute for Plasma Physics at the University of Nagoya, the Hitachi computer at the University of Tokyo, and the NEC computer at the Institute for Laser Electronics, University of Osaka. These "beta test" sites provided early exposure to a sophisticated scientific population. In each case the vendor provided extensive on-site support to the university computer center and received rapid feedback on user experiences and improvements. Because both the Fujitsu and Hitachi computers used IBM compatible front ends and executed the IBM instruction set for scalar operations, a large number of systems and applications codes were quickly adapted for use on these supercomputers.

User reports indicate that both the hardware and software of these computers are very reliable, roughly up to the level of Cray's offerings. To date there are fewer applications codes available, but sophisticated users tend to write their own codes to solve

problems at the frontier of science. A surprise to American observers has been the sophistication of the Fujitsu vectorizing FORTRAN compiler. All modern supercomputers employ "vector" processors to speed up performance on ordered data. Vector processors apply the same instruction to an ordered array, but it can be difficult for a compiler to recognize code in which vectorization is possible. The Fujitsu compiler is one of the most sophisticated available in this regard, and its sophistication effectively increases the speed of the Fujitsu VP series. The research on which the Fujitsu compiler is based was performed at the University of Illinois and Rice University with funding from the U.S. Government.

All three Japanese supercomputer vendors have increased their market share in semiconductor manufacturing, providing not only the chips needed for their own supercomputers, but also many of those used in American supercomputers. The Japanese supercomputer vendors are now at the leading edge of semiconductor technology and have the capability to design and fabricate specialized chips for their own use.

One area where the Japanese lag is in parallel processing. Parallel processing involves using several independent arithmetic processors to compute different elements of a problem simultaneously. This is the analog of a team of horses that works together to pull one carriage. U.S. supercomputers built by Cray Research have incorporated up to four processors since 1985 and will soon contain up to sixteen processors. The ETA-10 built by CDC/ETA has eight processors. To date, the Japanese have not announced supercomputers with parallel processors, although it is known that they are working to develop them. Parallel processing is much like vectorizing. Rethinking and recoding are required to take advantage of the theoretical performance improvement. Applications programmers have been learning for ten years how to vectorize code; possibly they may take that long again to learn how to parallelize code. To date, Cray has offered only rudimentary software tools to assist programmers in parallelizing code, and very few applications codes presently exploit parallel processing. In the near term, parallel processors are likely to be used as a collection of tightly coupled individual computers, with each processor running a different code. If this proves true, the Japanese can compete on the basis of single processor performance and cost. History suggests that this will be the case: when the Cray 1 was introduced, its immediate competitive advantage was not its ability to vectorize, but rather the fact that its scalar speed was twice as fast as its competition.

B. Current Status of Japanese Supercomputer Industry

Three Japanese companies currently offer supercomputers: Fujitsu, Hitachi, and NEC. All three companies are giants compared with Cray and CDC/ETA; each offers a broad line of computers and other electronic products. Both Fujitsu and NEC have recently begun to offer supercomputers for sale in the U.S., Fujitsu through Amdahl (of which it is part owner) and NEC through Honeywell. To date only one Japanese supercomputer has been installed in the U.S., a NEC SX-2 provided at a large discount from list price to the Houston Area Research Center (HARC). The SX-2 is the fastest single processor machine in the world, with speeds up to 1300 megaflops. Announcement of a multiprocessor version of the SX-2 is expected in 1988.

The Japanese supercomputer manufacturers gain a great advantage over their American counterparts by producing their own microcircuits. In fact, these companies are world leaders in manufacture of microcircuits, with NEC, Hitachi, and Fujitsu ranked number 1, number 2, and number 7 in the world, respectively, as semiconductor manufacturers.[18] Of even greater importance is the fact that, because they are world leaders in these microcircuits, they are the major source of supply for their American competitors. For example, virtually all of the memory chips and half of the logic chips in the Cray X-MP are purchased from Fujitsu, and there is no satisfactory American second-source for many of these circuits.

Effective control of Japanese companies rests more with the banks as stock holders as well as lending institutions than with individual investors. Because of the general Japanese business environment, Japanese manufacturers are under less pressure to increase near-term profits than their American counterparts and are able to focus more on long-term research and development. Generally, the debt/equity ratios of Japanese companies are much higher than for American companies.

C. Role of Japanese Government

As in other industries, the Japanese government plays several roles in the supercomputer industry: funding

source for R&D, leader in developing industrial and export policy, and provider of "beta test" sites at its universities and laboratories. Each of these roles has benefited the Japanese supercomputer industry in demonstrable ways.

Although the overall Government share of R&D in Japan is relatively small compared with the U.S. (only about 25 percent of total R&D), it is targeted towards areas considered to be critical to the Japanese economy.

MITI Involvement in Information-Related Technology Development

Appendix G contains a translation of an article entitled "MITI Involvement in Information Technology Development" which describes Japan's Government support of supercomputer R&D. A summary follows:

Summary

Information technology is developing into a leading Japanese industry. MITI support of the computer industry can be traced back to the Industrial Testing Grant in 1950. Until the 1980's MITI supported R&D was of a "catch up with the West" mentality. Now there is a shift to support of long-range high risk R&D that is deemed too expensive for the private sector. Overall, MITI tries to support research in fields that are critical to the Japanese economy. The Fifth Generation Computer Project is an example of the new emphasis. This 10 year project is intended to supplant the fourth generation VLSI computers with a revolutionary new computer. It began with surveys in 1979 and R&D work commencing in FY 1982. Current annual funding is about $130 million.

Information technology accounts for about 15 percent (about $150 million) of the MITI FY 1986 and FY 1987 budget. The level of support reflects MITI's belief that information technology is important for the future of the Japanese economy.

Other major MITI supported projects include an effort to build systems that can use a diverse variety of databases in text, graphics, video and audio (Computer Interoperable Database System) and a program to stimulate the development of the software that is needed to fully utilize new computer systems. (System for Industrializing Software Production-Sigma System).

A planned MITI supported activity is to promote international cooperation in order to counter criticism that Japan is getting a free ride by applying technology developed in the West. The approach is to undertake and aggressively publicize basic research projects.

The report states that the Japanese Government supports only 25 percent of technological development which is lower than in western nations. Increasing this percentage is an "urgent necessity," but the government's tight budget makes direct funding difficult. Hence, the Japanese private sector will be relied on for the bulk of technological progress. The specially licensed corporation called the Key Technology Center, which was formed with private financing in 1985 is seeking to engage private enterprise in basic research.

Strengthening supercomputing is such a priority, and the Ministry of International Trade and Industry (MITI) supports several research projects with industrial and university involvement. Examples of Government funded R&D related to supercomputing include the $100 million Super-Speed Computer Project funded by MITI with research performed by Fujitsu, Hitachi, NEC, and other Japanese companies, the Next-Generation Industries Project to develop advanced components needed by Japanese supercomputer firms, and the Fifth-Generation Computer System Project with annual funding of about $36 million. (See Appendix G.) By comparison, virtually the only industrial computing R&D funded by the U.S. Government is for defense purposes. The reduction of corporate risk by government funding and by cooperative R&D is considered by some students to be the major single Japanese competitive advantage.

The Japanese Government has targeted information technology generally, and supercomputing specifically, for national leadership and development of export markets. The Government (especially MITI) has been quite successful in its previous efforts to steer Japanese industry in preferred directions such as consumer electronics and semiconductors. For the supercomputer industry the Government has expedited export approval, which usually takes only a few days, and has protected the Government sector of the Japanese marketplace from American supercomputers.

National universities funded by Monbusho (The Ministry of Education, Science and Culture) provide

hospitable "beta test" sites (friendly users of new machines) for Japanese supercomputers, with each major vendor having a continuing relationship with a particular university: Hitachi with the University of Tokyo, Fujitsu with Nagoya University's Institute for Plasma Physics, and NEC with Osaka University's Institute for Laser Electronics. As each new supercomputer is developed it replaces the previous high-end model at that university. On site vendor personnel work with university scientists to correct bugs and improve performance before the computer is released into the commercial marketplace. Japanese supercomputer manufacturers give very large discounts to these university sites, ranging up to 80-90 percent off list prices.

D. Near-Term Projections of Japanese Capability

Recent progress in Japan's supercomputing industry plus analogies with other industries allow a near-term projection that should be fairly reliable. It has been made clear in public statements that supercomputing is a targeted industry and that Japanese vendors will employ very aggressive marketing tactics buttressed by their formidable economic strength and existing marketing arrangements with U.S. companies. Japanese leadership in components will probably increase because of the relative R&D expenditures and capital investments of Japanese and U.S. semiconductor manufacturers. A serious near-term threat would be the withholding of leading edge logic and memory technology by Fujitsu, Hitachi, and NEC to secure competitive advantage for their supercomputers. R&D successes may allow Japanese companies to bring new semiconductor technologies to the supercomputer. For example, although the Cray 3 may be the first supercomputer to use high speed gallium arsenide chips, the Japanese are devoting more R&D to develop gallium arsenide and related high speed semiconductors. They have also maintained significant R&D in superconducting Josephson junction gates, whereas until recently the U.S. has largely abandoned this field. The recent discovery of high temperature superconductors could make very high speed, low power, supercomputers based on Josephson junction technology possible, operating at liquid nitrogen temperatures. Since ETA has already demonstrated the use of liquid nitrogen to cool a silicon based supercomputer, the Japanese might be able to develop a Josephson junction computer in a relatively few years.

VI. Findings and Recommendations

A. Findings

A vigorous domestic supercomputer industry is essential for maintaining U.S. leadership in critical defense areas and in areas important for our civilian economy.

U.S. leadership in many critical technology areas has been based on leadership in developing and exploiting supercomputers. This leadership would be jeopardized by dependency upon other countries for supercomputers.

Supercomputers play a vital role in R&D of weapons systems. Emerging parallel processors are likely to be the nerve centers for the Strategic Defense Initiative. Loss of a strong domestic supercomputer industry would make us dependent upon other countries to develop and supply these critical tools for our defense.

Supercomputers have developed into a vital part of our science and technology infrastructure. Computer simulation is approaching equal footing with theory and experiment in research. The potential for faster supercomputers to contribute to the basic research funded by the Federal government is enormous.

The growing importance for industrial application makes supercomputer leadership vital for industrial competitiveness. Government use of supercomputers has spawned industrial applications that confer competitive advantage to the user. Industrial applications now constitute more than half of the supercomputer market and contribute importantly to U.S. industrial competitiveness. This is partly a consequence of effective technology transfer from the federal laboratories that use supercomputers.

U.S. supercomputer leadership is threatened

The U.S. currently leads the world in research, development and use of supercomputers. However, at

present, the US supercomputer industry consists of only two small companies, with other small companies developing advanced architectures that may become tomorrow's supercomputers. It will be difficult for these companies to continue to compete successfully against large vertically integrated Japanese companies. The Japanese Government has targeted information technology and supercomputers as national growth areas, and MITI is funding cooperative R&D projects with Japanese manufacturers. Japanese markets are difficult for U.S. supercomputer manufacturers to penetrate, while American markets remain open. The American supercomputer manufacturers are dependent on their Japanese supercomputer competitors for essential semiconductor components, who are assuming world leadership in integrated circuits.

The Federal Government has retreated from its historic role as "friendly" buyer of supercomputers.

The Federal government first recognized its need for high performance computers during World War II. In the 1950's and 1960's weapons design and other defense needs, as well as the growth of government support of basic research, made the Federal government the major purchaser of supercomputers. At this time federal policy allowed the purchase of innovative new machines that were sold with very limited software. Talented individuals at the laboratories, principally Lawrence Livermore and Los Alamos National Laboratories, developed the necessary software and provided initial operating experience. Much of this software then became generally available for other purchasers of the machines and the operating experience helped less sophisticated customers to determine the value of the new computers. This "friendly" buyer support has successfully spawned a plethora of private sector applications of super-computers. Now industry buys more than half of the new supercomputers.

Procurement constraints and the growth of private sector demand for supercomputers have contributed to the the Federal governments retreat from its role of "friendly" buyer. However, supercomputers have continued to grow in importance to national defense programs. The emergence of complex new supercomputer architectures presents additional challenges to their effective application. Loss of the ability to test these new computers in sophisticated environments slows the pace of development and confers a competitive advantage to Japanese suppliers with their university test sites.

U.S. leadership in parallel processing is a key ingredient for maintaining supercomputer leadership

For the past three decades, large scale scientific computers have been designed around a central processor that performed operations sequentially to produce the desired result. This computer architecture has witnessed a tenthousand fold increase in speed since 1960. However, solid state electronic components are now approaching fundamental physical limits of the speed that they can provide for the conventional sequential processor approach.

Multi-processor machines can perform more than one operation at the same time, i.e., in parallel. At present, parallel processing architecture is the most promising approach to producing significantly faster supercomputers. The U.S. is now the leader in the development of parallel processing hardware and software. Exploiting parallelism effectively presents formidable challenges. If properly employed, U.S. skill and experience in these areas can help to preserve national supercomputing leadership.

Needs of the supercomputer industry must be considered in formulating trade policy.

Some aspects of U.S. trade policy have had unintended negative impacts on the U.S. supercomputer industry. The recent U.S./Japan semiconductor chip agreement has resulted in increases in prices of some chips for which the supercomputer industry has no other source. It appears that policy makers completed the agreement without seeking the views of the US supercomputer industry. This has been counterproductive to our national trade objectives, because the agreement made it more difficult for U.S. companies to compete in certain product areas.

National security controls placed upon the export of supercomputers and related products have placed U.S. companies at a competitive disadvantage with respect to foreign manufacturers in certain markets. Typically U.S. supercomputer manufacturers experience much longer delays in obtaining export licenses than their Japanese competitors.

B. Recommendations

The U.S. Government should provide an environment favorable for a vigorous domestic supercomputer industry.

Supercomputers have become an essential tool in scientific and engineering research. This research, in turn, supports our technological edge in national security and industrial competitiveness. Supercomputers are vital for weapons design, intelligence analysis, weather modeling, and many other areas of recognized national interest. Key industries such as petroleum, automotive, electronics, chemical and aerospace are finding supercomputers essential for remaining competitive. Numerical experimentation, using ever more powerful supercomputers, is developing into a complement to the theoretical and experimental methods which have been the cornerstone of modern technological progress.

Japan has been strengthening its national commitment to supercomputer development since 1981. The large, vertically integrated Japanese supercomputer manufacturers are making great strides toward overtaking the two, relatively small, U.S. manufacturers. American preeminence in supercomputers is clearly threatened.

It is inconceivable for the US to accept dependence on another country for computers that are so critical for national defense and economic competitiveness.

The Federal Government should return to its role as a "friendly" buyer of the latest supercomputers.

- **Government agencies should budget for acquisition by their laboratories and contractors of prototype or "serial one" models of new supercomputers that offer potential for improving their research productivity.**
- **These initial acquisitions should not require complete operating systems and applications software typical of production computers.**

Since the 1950's, the Federal government has purchased the first, or one of the first of each new large scale computer system with minimal operating software and no applications software. Profits from the sales of one generation of computers financed R&D for the next generation. The laboratories, particularly Lawrence Livermore and Los Alamos National Laboratories, developed substantial amounts of software required to make the machines into useful systems. This "friendly" buyer role stimulated the private sector to develop even faster supercomputers that have maintained in U.S. preeminence. The current market for supercomputer is small and development costs are large. At present, there are only two US manufacturers of supercomputers: Cray Research Inc. and ETA Systems Inc.

The Federal government still accounts for nearly half of all supercomputer sales. Current government programs require supercomputer capability far in excess of what is available today. Within this decade programs in nuclear weapons design, aerospace, weather prediction, fusion research and many areas of fundamental R&D will require computers of capability hundreds of times greater than what is available today. The development of the advanced designs is too risky for the small US industry to undertake without assurance that their new machines will be purchased.

The Federal Government should make federally developed software when possible available to the private sector.

- **Manufacturers should be expected to develop software for production computers, but software developed by government laboratories for prototype computers should be transferred to the private sector when national security concerns allow.**

The U.S. Government should carry out a coordinated long range R&D program in supercomputer applications, software, and investigation of advanced computer architectures.

- **A primary objective is to ensure continued private sector development of supercomputers that are required for use in Federal programs. (Current national defense and basic science and engineering and programs could benefit significantly from faster supercomputers.) The program should also be based on an anticipation of future needs. The program will have the beneficial by-product of aiding the development of supercomputers for other industrial applications.**
- **The program should build upon existing government-supported efforts.**
- **Major scientific and engineering challenges should serve as a focus for effort.**

- **Advanced computer architectures should be developed.**
- **A 1000 fold improvement in applied computational capability in five years should be a short-term goal.**
- **R&D should address algorithms, both systems and applications software, and peripherals for future supercomputers.**

A long range research program is justified by the importance of US supercomputer leadership for national defense and industrial competitiveness. The requirements of government programs result in the Federal purchase of nearly half of all supercomputers. The Federal government remains a major user of supercomputers.

The needed research is long range and high risk that has potential for a high return in the future. The US industry, which consists at present of only two small companies, is not capable of funding the research needed to keep ahead of their large Japanese competitors which have priority Japanese government support.

The recommended program should provide a coordinated enhancement of existing Federal programs. Emphasis would be on supporting private sector R&D.

There are numerous scientific and engineering problems in existing Federal programs that could provide grand challenges that would focus federally supported research programs. Supercomputers enable researchers to tackle problems that were previously too complex and time consuming to attempt. Weather prediction, climate modeling, ocean circulation and the build up of carbon dioxide in the atmosphere are examples of areas where the pace of progress is dictated by the speed of available computers. Industry has rapidly applied the computers developed for Federal programs to a diverse array of private sector application.

Because the speed of electronic components are beginning to approach fundamental limitations, the major improvements in computational speed will probably come from different architecture. The United States currently leads the world in parallel processing, the most promising new computer architecture. However, Japan has identified this area for a focused effort. The development of novel architectures represents a high risk, high pay off effort for which the small US industry lacks adequate financial resources.

A 1000 fold improvement in applied computational capability represents an achievable goal that is consist with the rate of progress in the past.

An investment in algorithm and software development is an integral part of any large scale computational program. Federal laboratories represent a valuable resource of technical expertise in software and algorithm development.

Acknowledgement

The Committee would like to acknowledge the efforts of Dr. Michael D. Crisp who edited and contributed to the report.

References

1. Report of the Panel on Large Scale Computing in Science and Engineering, Peter D. Lax, Chairman, December 26, 1982.
2. Report to the Federal Coordinating Council on Science, Engineering and Technology (FCCSET) Supercomputer Panel on Recommended Government Actions to Retain U.S. Leadership in Supercomputing (January 1983).
3. Report to the Federal Coordinating Council on Science, Engineering and Technology Supercomputer Panel on Recommended Action to Provide Access to Supercomputers (January 1983).
4. A National Computing Environment for Academic Research, NSF, July 1983, M. Bardon and K. Curtis.
5. Report of IEEE United States Activities Board Scientific Supercomputer Committee Chaired by Sidney Fernbach (October 25, 1983).
6. "Strengthening the National Supercomputer Capability," Institute for Defense Analysis Panel Chaired by Jacob Schwartz (May 31, 1985).
7. Report of the Federal Coordinating Council on Science, Engineering and Technology Panel on Advanced Computer Research in the Federal Government (June 1985).

8. Report on Interagency Networking for Research Programs by the FCCSET Committee on Very High Performance Computing (February 19, 1986).

9. "Computers and Their Role in Energy Research Current Status and Future Needs." Hearing before the House Science and Technology Committees' Subcommittee on Energy Research and Production, June 14 and 15, 1983.

10. "Federal Supercomputer Programs and Policies." Hearing before the House Science and Technology Committee's Subcommittee on Science, Research and Technology, June 10, 1985.

11. "Targeting the Computer" Kenneth Flamn, The Brookings Institution.

12. Federal Coordinating Council on Science, Engineering and Technology (FCCSET) Committee on High Performance Computing Annual Report, Jan. 1987.

13. U.S. Congress, Office of Technology Assessment, *Supercomputers: Government Plans & Policies—A Background Paper,* OTA-BP-CIT-31 (Washington, DC: U.S. Government Printing Office, March 1986).

14. H. J. Ravaché et.al., *Report of the Panel on Applications of High Performance Computing in Engineering and Science,* Society of Industrial and Applied Mathematics, 1987.

15. Sanford C. Bernstein & Co., Projections of Supercomputer Marketplace, 1985.

16. Electronics News, March 30, 1987.

17. *Balancing the National Interest* US National Security Export Controls and Global Economic Competition; National Academy Press (1987).

18. Unpublished Report, Dataquest Inc., Dataquest's Semiconductor Industry Group June, 1987.

FCCSET COMMITTEE ON COMPUTER RESEARCH AND APPLICATIONS

Paul G. Huray, Chair
Office of Science and Technology Policy

SUBCOMMITTEE ON SCIENCE AND ENGINEERING COMPUTING

James F. Decker, Chair
Department of Energy

James Burrows
National Bureau of Standards

John S. Cavallini
Health and Human Services

Melvyn Ciment
National Science Foundation

John Connolly
National Science Foundation

Craig Fields
Defense Advanced Research Projects Agency

Harlow Freitag
Supercomputer Research Center

Randolph Graves
National Aeronautics and Space Administration

Norman H. Kreisman
Department of Energy

Lewis Lipkin
National Institutes of Health

Allan T. Mense
Strategic Defense Initiative Office

David B. Nelson
Department of Energy

C.E. Oliver
Air Force Weapons Lab

John P. Riganati
Supercomputer Research Center

Paul B. Schneck
Supercomputer Research Center

K. Speierman
National Security Agency

Appendicies

APPENDIX A

AN UPDATE ON COMPETITION IN THE SUPERCOMPUTER INDUSTRY: JAPAN vs USA

by
Jack Worlton
Los Alamos National Laboratory

Submitted to:

The Federal Coordinating Council
on
Science, Engineering, and Technology

Washington, D.C.

February 28, 1987

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

Purpose, scope, and method. This report is being prepared at the request of Dr. James Decker, on behalf of the Federal Coordinating Council on Science, Engineering, and Technology (FCCSET). The deadline for the report is February 28, 1987, a month in which the author is on travel for 20 of the 28 days, so this report is necessarily brief and incomplete (including incomplete editing). The method employed in this report is that of a situation audit, which is conceptually a matrix: on the left side are strengths plus resulting opportunities and weaknessess plus resulting problems; on the top are the relevant firms from the U.S. (Cray Research, Inc. (CRI), and CDC/ETA Systems) and Japan (Fujitsu, Hitachi, and NEC). We have not included IBM in this report, because their position is somewhat ambiguous: the IBM 3090VF is not considered a supercomputer by IBM for purposes of avoiding export controls, but it is considered a supercomputer by IBM when certain customers are being approached. In the long term, IBM should probably be included in this kind of analysis.

Background. In 1983 two Japanese firms (Fujitsu and Hitachi) began early deliveries of supercomputers in Japan, and in 1985 NEC began deliveries of their supercomputer. Prior to this time, the supercomputer market had been a small and exclusively American market, with only Cray Research and CDC offering such computers. In 1983 CDC formed a new small company, ETA, to develop their next generation of supercomputers; CDC is the primary stockholder in ETA, so this company is sometimes referred to as CDC/ETA or simply as ETA. The American and Japanese firms are quite different: CRI has revenues of a few hundred million dollars per year and ETA is still just a startup company, whereas the Japanese firms have annual revenues on the order of $10 to $20 billion. The American firms are not semiconductor manufacturers and must depend on other companies for their components, whereas the Japanese firms are all world leaders in merchant semiconductor manufacturing. On the other hand, because the American firms were in this market before the Japanese firms, they had a market share advantage, so the problem for the American firms was to hold their market share, whereas the problem for the Japanese was a matter of penetrating an existing market. The Japanese firms enjoy several structural advantages, the most important of which is the "partnership economy" of Japan, in which the government and industries of Japan work as partners in promoting industrial expansion. In the U.S., the government and industry are often adversaries, with the most obvious example being export controls which constitute a major problem for the American firms but not for the Japanese firms. Since their entry into this market in 1983, the Japanese firms have been most successful in Japan, with Fujitsu leading in installing computers there, at last report about 30. None of the Japanese firms has so far been particularly successful in the international market; Fujitsu has been the most aggressive, but NEC (whose motto is *Attack*!) is now supplanting Fujitsu as the most aggressive

in international supercomputer marketing. Hitachi has limited their supercomputer marketing to Japan.

Situation audit. The reasons for the ability of the American firms to hold their market share in the face of competition from such strong Japanese firms include (1) market share, (2) technical leadership, and (3) an existing marketing infrastructure. All other things being equal, a large market share tends to be self-perpetuating because of repeat orders, compatibility considerations, the desire of customers to collabore with other sites having the same type of computers, and a large and stable base of system and application software. The task facing the Japanese firms is to attack the phrase "all other things being equal." The strengths of the Japanse firms that might change this assumption include their strong semiconductor development capabilities, which can be used to develop advanced components one to two years before the American firms, and their financial strengths, which can be used to "buy" contracts away from the smaller American firms who cannot afford the heavy discounts being offered by the Japanese firms. If this competition were being conducted wholly within the U.S., antitrust laws would prevent the Japanese firms from using many of their marketing strategies such as "anticipatory pricing" (selling below cost), but in the international market, no such rational protections exist.

Specifically, the strengths of Cray Research include a market share of about two-thirds of current supercomputer installations, technical leadership in parallel processing which they began in 1982, a rich base of application and system software, and a strong marketing and technical support infrastructure. Their primary weakness is their dependence on other firms to provide the high-performance logic and memory components they need for new generations of supercomputers. CDC/ETA has a smaller, but not inconsiderable market share of some thirty-odd machines, which will be an advantage to them in marketing their new machine, the ETA^{10}. They have long experience in developing supercomputers and hence a strong and knowledgeable staff. Their weaknesses include the fact that they have no current product to sell, with the ETA^{10} hardware and software still being in development and their Cyber 205 being obsolescent; their status as a startup company with little or no income; and their dependence on other firms for advanced components.

The specific strengths of the Japanese firms include their leadership in semiconductor manufacturing, which should give them a timing advantage in developing new generations of supercomputers; the support they receive from the Government of Japan in the form of government-supported research projects and avoidance of export licensing problems; and their financial strengths which can be used to support deep cuts in pricing (referred to as "anticipatory pricing"). Their weaknesses include a small market share, immature software, incompatibility with most supercomputer users, lack of credibility for software and maintenance support, and architectural obsolescence (marketing serial processors in a world rapidly moving toward parallel processing).

Summary. The "first round" of international competition in supercomputing must be conceded to the American firms, primarily Cray Research. Japanese successes have been limited largely to Japan where cultural preferences have made it easy for the Japanese firms to corner that market. There are widely circulated rumors of a new generation of supercomputers from the Japanese firms being introducted in 1987, although no formal announcements have been made. Both Cray Research and ETA Systems also plan to offer new computers in 1987, so it is possible the "second round" will still be a wash, unless there are some big surprises from the Japanese firms. It is in the long term that the Japanese advantages of component development, financial strengths, and government relations will be most evident. To survive, the American firms must somehow gain access to timely development of high-performance logic and memory components independent of Japanese firms, and it is not yet clear how they will do that. Collaboration with some small "niche vendors" seems to offer the best hope at the moment. The American firms also need to somehow counter the Japanese advantage in government relations. Whether this should take the form of direct government support as in Japan, or merely removing governmental barriers such as export control delays, is a topic on which there is agreement only on the latter point.

2. INTRODUCTION

2.1 Purpose. The purpose of this report is to conduct a brief "situation audit" of competition between Japan and the USA in the supercomputer industry. Prior to 1983, the supercomputer market had been exclusively American for about twenty years, with the last supercomputer marketed by a foreign country being the British ICL Atlas from the early 1960s. However, in 1983, two Japanese companies, Fujitsu and Hitachi, began deliveries in Japan of supercomputers whose performances were within about a factor of 2 of the performance of the leading American supercomputers, and in 1985, NEC began deliveries of computers that were faster than the Fujitsu and Hitachi machines by about a factor of 2. As shown in Table 1, these products--the so-called "Developing Generation"--were the result of earlier developments--the so-called "Embryonic Generation"--by these companies."

	EMBRYONIC GENERATION	DEVELOPING GENERATION
FUJITSU	◦ FACOM 230-75 APU -- 1977 -- 22 MFLOPS -- Shared main memory -- Fully pipelined -- 1972 vector registers -- AP Fortran	• VP-100/200/400 -- 1983 -- 250/500/1000 MFLOPS (peak)
HITACHI	◦ M-180 and M-200 w/IAP -- 1978 and 1980 -- 24 MFLOPS -- Vectorizing compiler -- Shares func. units w/CPU	• S810/10 and 20 -- 1983 -- 315/630 MFLOPS (peak)
NEC	◦ ACOS 1000 W/IAP --1981 -- ?? MFLOPS	• SX-1 and SX/2 -- 1983 -- 570/1300 MFLOPS (peak)

Table 1. Embryonic and Developing Generations of Japanese Supercomputers.

The "IAP" refers to an "Integrated Array Processor" that was an arithmetic accelerator attached to the mainframes. Thus, these companies did not, as often believed, suddenly begin producing supercomputers, but had been working on vectorizing units and their software for several years. Both of these generations borrowed ideas from prior American designs, including

the CDC Star-100, the Cray-1 and the CDC Cyber 205 Since the introductions of the Japanese products in 1983 and 1985, several products have been added on the low end of the cost and performance range.

It is anticipated that the third generation of Japanese supercomputers--presumably a "Mature Generation"--will be forthcoming in the next year or so. Very little information but lots of rumors have been circulated concerning these machines.

2.2 Scope. The scope of this report is severely limited by the time constraints allowed for its preparation, so it is mostly an incomplete digest rather than a complete and detailed report. Further information can be obtained by checking the sources noted in Section 7, References and Bibliography.

2.3 Method. This report is a situation audit, with emphasis on key issues facing the competitors and their host countries. We include not only the usual marketing and technical issues, but also the "structural" issues that are crucial for understanding any competitive situation with respect to Japan. This is often referred to as "The Japan Problem."

2.4 Taxonomies. To clarify the class of computers under discussion, we include two taxonomies. Figure 1 is a partial taxonomy of high-performance scientific computers that shows the three main categories of such machines: research, special-purpose, and general-purpose. We shall be concerned here with general-purpose high-performance computers. Within that category are three types of computers: supercomputers, high-end mainframes, and "mini-supers." Although there is some overlap in the performance ranges of these types of computers, the supercomputers as a class outperform the other two types of high-performance computers and this category is the subject of this report.

Within the category of supercomputers, there are three "classes" often referred to, as shown in Figure 2. The performance ranges shown are only approximate, of course. Supercomputers by their nature have very broad performance ranges compared to other kinds of computers, and the overlap of the performance ranges for the three classes is deliberate. Whereas the first two classes of supercomputers represent machines that have already been delivered (or are reasonably close), none of the Class 7 machines have been delivered, and these are merely announced plans of the companies whose products are shown.. The new generation of Japanese supercomputers will presumably fit into the Class 7 category.

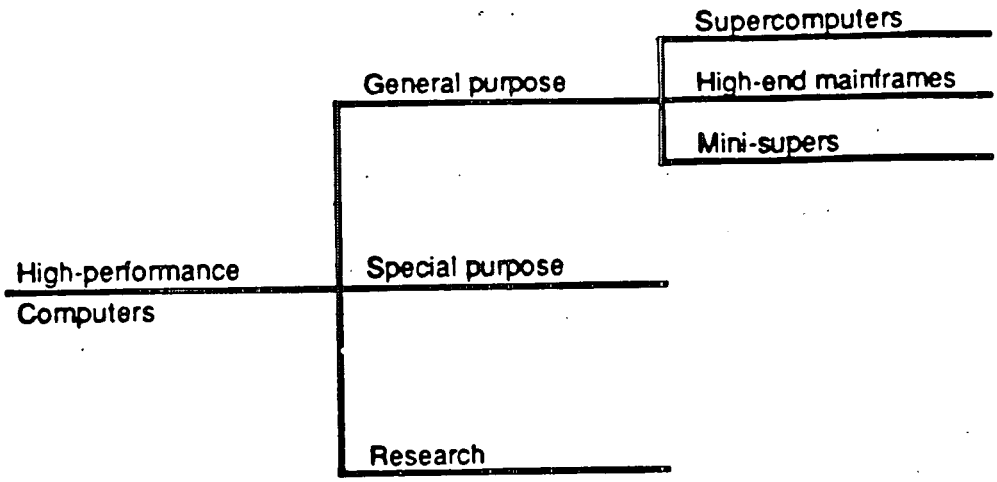


Figure 1. Partial Taxonomy of High-Performance Computers.

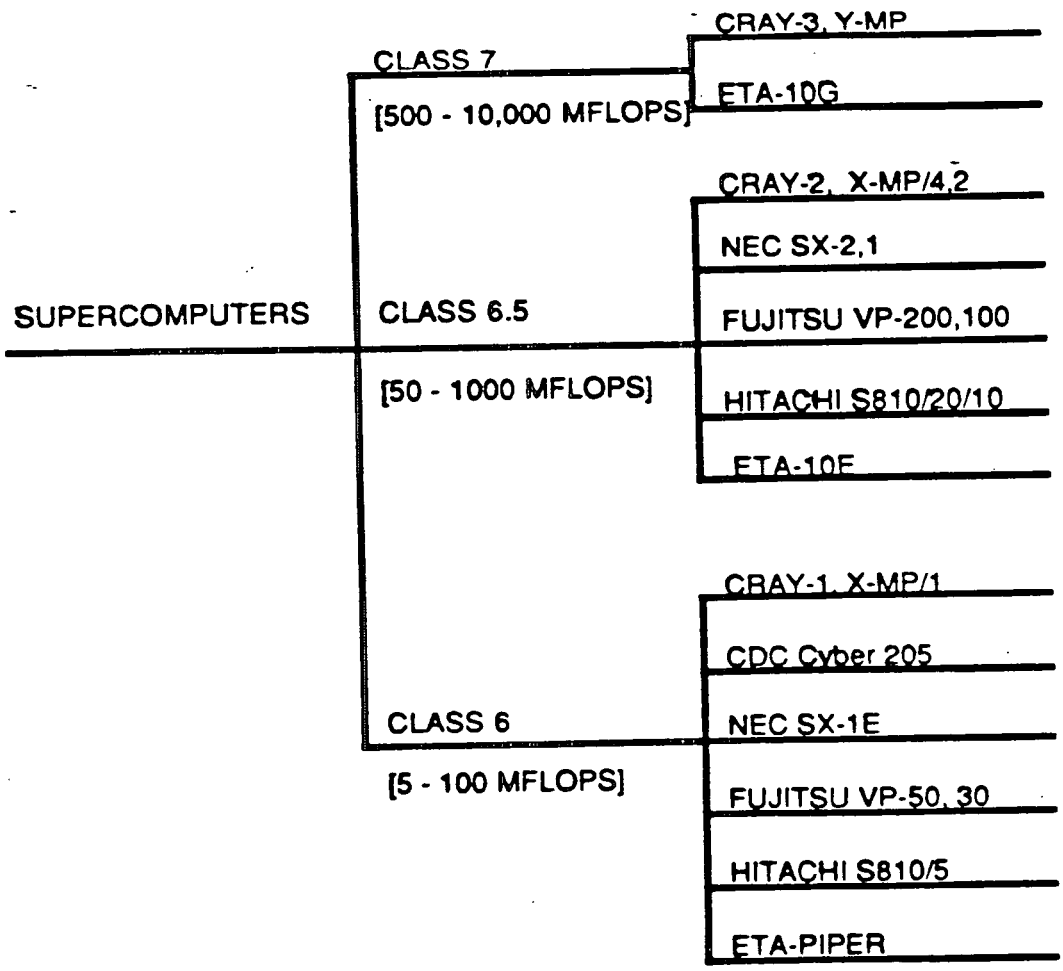


Figure 2. Taxonomy of supercomputers.

3. MARKETING ISSUES

3.1 Market share. Figure 3 shows the share of the supercomputer market held by Cray Research, CDC/ETA, and the Japanese companies as of mid-1986.

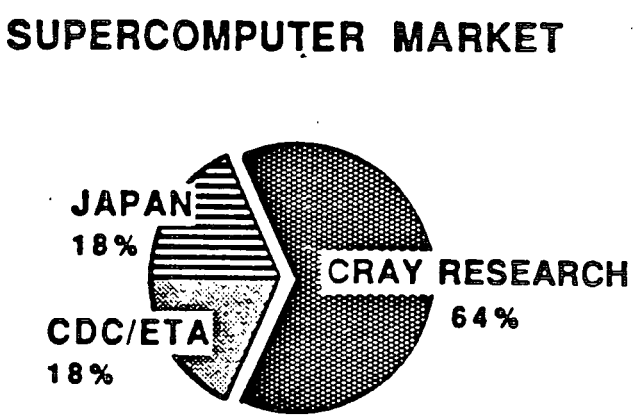


Figure 3. Shares of the supercomputer market as of mid-1986.

The percentages shown here are continually changing, of course, but it is roughly true that Cray Research sells about two-thirds of the supercomputers in the world, and the other third is divided between the three Japanese companies and CDC/ETA. Or, to put it another way, since the Japanese began delivering supercomputers in 1983, they have captured roughly 18 percent of the market. However, a basic principle of marketing is that *a large market share tends to be self-perpetuating*, because of the commitment of the customer to the particular product, where in this context "commitment" includes such things as applications codes, user competence and training, operational skills, and site installation. These commitments are reflected in repeat orders, a rich body of application and system software, and collaboration with other sites using compatible computers. And while this is true of computers in general it is true of supercomputers in particular. The reasons for this lie in the effort needed to prepare applications software for these computers and the rich set of software that the customers can obtain with a minimum effort or expense. More software exists for the computers having large market shares because there are more users and more third-party software vendors developing such software for these computers. This is true of IBM's large share of the mainframe market and Digital Equipment Corporation's share of the minicomputer market, as well as Cray Reasearch's share of the supercomputer market. Thus, Cray Research's large market share is one reason the Japanese companies have not made more progress than they have in penetrating the supercomputer market.

3.2 Timing. Another reason Cray Research has been able to hold off this intense competition from powerful Japanese companies is found in the timing of recent introductions. Cray Research

introduced the Cray-1 in 1976, and this machine had essentially no competition for about five years and only minimal competition for some after that until the internal competition created by the introduction of the Cray X-MP/2 in 1983. The Japanese vendors targeted the Cray-1 as the computer their computers should exceed in performance. However, by the time Fujitsu and Hitachi entered this market in 1983, Cray Research had introduced in 1982 a newer and more powerful product line, the Cray X-MP/1 and Cray X-MP/2, with one and two processors, respectively. The single-processor X-MP has about the same performance as the Fujitsu VP-200 in general-purpose computing, and about twice the performance of the Hitachi S810/20. Thus, these Japanese computers were indeed faster than the Cray-1, but this was no longer the relevant comparand by the time the Japanese products were introduced.

A similar situation occurred when NEC introduced the SX-2 in 1985. This computer was faster than the other Japanese supercomputers by about a factor of 2. However, by the time NEC introduced this computer Cray Research had introduced both the four-processor X-MP/4 and the four-processor Cray-2, thereby effectively preempting the NEC introduction.

During this period, Control Data Corporation, the vendor of the Cyber 205 (a computer roughly in the same performance category as the Cray-1) formed a new subsidiary, ETA, to design the next generation of supercomputers, generically referred to as the ETA[10]. These computers are in the late stages of their development, and ETA is expected to begin deliveries of hardware and software for the ETA[10] in the next twelve to eighteen months.

3.3 Compatibility. A commonly used guideline in the management of scientific computing is that *an incompatible computer must provide a performance gain commensurate with the cost of conversion*. This is usually quantifed as a factor of 2, i.e., an incompatible computer must outperform a compatible computer by at least a factor of 2 to justify the cost of conversion. The Japanese computers failed to meet that criterion even with respect to a single-processor X-MP, let alone the dual-processor X-MP. Thus, it was not surprising that neither the Fujitsu nor the Hitachi products were able to penetrate this market except in Japan, where the well-known Japanese antipathy toward foreign products led to the acquisition of mostly Japanese supercomputers in spite of normal computer evaluation criteria.

3.4 Market strategies. The Japanese employ two distinctive strategies when attempting to penetrate a new market: targeting and anticipatory pricing. Targeting is a national industrial strategy and anticipatory pricing is a corporate strategy. Targeting refers to the practice of bringing overwhelming national resources to bear against a specific industry of another nation, such that the target industry is at a disadvantage. For example, the Japanese Ministry of International Trade and Industry's Super-Speed Computer System Project, for the period 1981 to 1989 and funded with $100 million from the Japanese national budget, brought together the resources of Japan's six largest computer companies in a national project to develop the supercomputer technology that

would allow Japanese companies to become world leaders in this field. The American companies in the supercomputing industry, Cray Research and ETA, have revenues of less than 1/10 to 1/40 of the revenues of the leading Japanese companies, and the aggregate resources of the six major Japanese companies are even more overwhelming.

The financial strengths of the Japanese companies make it possible for them to employ the second of these strategies, "anticipatory pricing." Essentially this means that a company sells its products either below cost or at huge discounts, attempting thereby to attract customers away from the companies whose products are priced to make a normal profit; this is sometimes referred to as "dumping." The word "anticipatory" refers to the expectation that in the long term, as the Japanese company gains market share, their prices will be adjusted to generate profits. This strategy cannot be employed by small companies that must make profits in order to survive, but only by large companies with other divisions whose profits support this penetration of a new market. The anticipatory pricing strategy is being used currently by Japanese firms offering huge discounts, a case in point being the well-known sale of NEC's SX-2 to the Houston Area Research Council (HARC) [4,5]; similar efforts are occurring in other nations. Briefly, this strategy is an attempt on the part of a large company to "buy" a market away from a small company and thereby put the small company out of business. NEC's company motto of *Attack!* is well illustrated by this strategy.

3.5 Architectural issues. The Japanese supercomputers are all single-processor designs, and this has probably had some negative effect on their marketing efforts. There is a broad consensus among the world's computer scientists that computers of the future, and especially supercomputers, will be built using multiple processors, so acquiring one of the Japanese machines has meant a customer was buying "instant obsolescence" in the architectural sense. Not many customers want to spend the millions of dollars supercomputers cost without getting a current design.

3.6 Product comparisons. Table 2 lists current and projected supercomputer products as of February 1, 1987. A few of the "Next Generation" products may be shipped in 1987, but substantial customer shipments are not expected until 1988. This is also true of the "Future Generation" for the years 1988 and 1989. Table 3 shows some general characteristics of some representative supercomputers from the current generation .

VENDOR	CURRENT GENERATION (Early 1987)	IN DEVELOPMENT (1987-88)	FUTURE GENERATIONS (1989 or beyond)
Cray Research	Cray-1 Cray X-MP/1,2,4 Cray-2	Cray Y-MP/8 Cray-3	Cray MP
CDC/ETA	Cyber 205	ETA-10G ETA-10E ETA Piper	Unknown
Fujitsu	VP-30,50,100 VP-200,400	Unknown	Unknown
Hitachi	S810/5,10,20	Unknown	Unknown
NEC	SX-1E,1,2	Unknown	Unknown

Table 2. Supercomputer generations.

SYSTEM	FCS*	Cycle Time (ns)	No. PEs	Main Memory (MW)	Extended Memory (MW)
Cray-1	1976	12.5	1	1-4	----
Cray X-MP/1,2,4	1982,1984	9.5/8.5	1,2,4	1-16	32-512
Cray-2	1985	4.1	4	256	----
CDC Cyber 205	1981	20.0	1	4-16	----
Fujitsu VP-200	1983	14/7	1	8-32	----
Hitachi S810/20	1983	14	1	4-32	32-128
NEC SX-2	1985	6	1	16-32	16-256

*FCS = First Customer Shipment

Table 3. General characteristics of some representative current-generation supercomputers.

3.7 Summary of marketing issues. In summary, Cray Research has been able to withstand the attacks of the larger Japanese companies during the past three years by virtue of its large market share, by its timely introduction of new products, by leadership in parallel processing, and by the incompatibility problem faced by the Japanese. ETA Systems is just now in the final phases of

product development, and it will about twelve to eighteen months before it is known how well they will do against their competitors, both domestic and foreign.

The continuing marketing problems facing the American competitors include (1) their limited ability to match the Japanese semiconductor-development capability, (2) the targeting strategy employed by the partnership between Japanese industry and the Government of Japan, and (3) the marketing strategies of anticipatory pricing and dumping employed by the much larger Japanese firms.

4. TECHNICAL ISSUES

4.1 Component technology.

The three Japanese supercomputer vendors, NEC, Hitachi, and Fujitsu, rank number 1, number 2, and number 7 in the world, respectively, as merchant semiconductor manufacturers [31]. The advantage this gives them in developing supercomputers is largely one of timing. They can develop new generations of advanced components about one to two years ahead of their American competitors, according to Tony Vacca, Vice President for Technology at ETA Systems. Suppose, for example, that a Japanese firm begins marketing a supercomputer with a 1 nanosecond (ns) cycle time two years ahead of a similar product from American firms. During this timing gap, there will be some market penetration before the American firms catch up, and after a few rounds of this experience, the total market would inevitably be captured by the Japanese. At the moment the fastest cycle times are found in the Cray-2 (4 ns) and the NEC SX-2 (6 ns). However, the Cray-2 issues instructions only every other cycle, so for scalar work its cycle time is more like 8 ns and only in long vectors does it appear as a 4 ns cycle. Thus, for practical purposes the Cray-2 cycle time is probably best thought of as about 6 ns. There are prospects for improving cycle times to 3 ns in silicon, 2 ns in gallium arsenide , and 1 ns in HEMT (high electron mobility transistors). For the American firms the problem is that the high-performance semiconductor market is so small that it attracts little attention from the major American semiconductor firms. Some small "niche" vendors who are willing to take a somewhat larger share of a small market have recently shown interest in serving this need, but they are evidently somewhat behind the Japanese firms in developing production versions of these advanced components.

4.2 Trends in supercomputer architecture.

There is an international consensus among computer scientists that the future of supercomputing lies with parallel processing, i.e., designs having multiple processors that can be used to shorten the solution time for a single problem. Cray Research began deliveries of such designs in 1982 with their Cray X-MP/2 with two processors, and in 1985 with their Cray X-MP/4 and Cray-2, each with four processors. The Cray Y-MP/8 will have a 5 ns or 6 ns clock and eight processors, planned for initial delivery in 1987; the Cray-3 is projected to have a 2 ns clock and 16 processors, for delivery in 1988 or 1989. ETA has parallel processors under development: the ETA^{10}-G with a 7 ns clock and up to 8 processors; the ETA^{10}-E with a 10.5 ns clock and up to 4 processors; and the ETA Piper with a 21 ns clock and 1 or 2 processors. There is, of course, the usual uncertainty about when systems under development will actually be delivered to customers in substantial quantities.

None of the Japanese firms has yet announced a parallel processor, although it is known that all of them are doing research on this kind of design, including the work being done on the Super-

Speed Computer System [28] sponsored by the Government of Japan. There have been some reports that this project is having problems, however [17]. In this sense, the Japanese supercomputers are architecturally obsolete, and this may have hindered some customers from taking an interest in these computers.

5. STRUCTURAL ISSUES

By "structural" issues we refer to those constraints on the supercomputer market that are built into the competitive environment over which the individual companies have little or no control, including the general trade relationships between Japan and the USA, cultural differences, and relationships between industry and government in the two nations.

5.1 The "Japan Problem."

An excellent summary of this problem is contained in a recent article by Karel G. van Wolferen, a Dutch writer who has lived in Japan since 1962 [1]. One of the myths about Japan is that it is a sovereign state like others among the Western nations. In other nations there is a source of power that can take responsibility for decisions and actions, but this is not true in Japan. Rather, there are *three* sources of power, none of which can assert that "the buck stops here," as did the American president Harry Truman. In Japan, the buck doesn't stop, it circulates among the politicians, the bureaucracy, and the industrialists. The evidence of this can be clearly seen in Nakasone's largely unsuccessful efforts to bring changes to the Japanese relationships with other nations. Western negotiators often express frustration at the seeming insincerity of Japanese negotiators because they say one thing during negotiations but do not follow through on apparent agreements when they get home. The problem here is caused not by Japanese insincerity but by a lack of understanding on the part of Westerners about the Japanese culture. Decision making on major issues can occur only by consensus among the three major centers of power, and thus it is impossible for any single party to represent the views of all three until later in time when a consensus has been reached. The point here is that Westerners should understand that Japanese politicians cannot make commitments in the same way politicians do in other Western governments.

A second myth about Japan is that its economy is market driven as are those of the other Western nations. Japan's economy falls into neither the free-market category nor the centrally-controlled category, but into what might be called a "partnership economy," where the partnership referred to is between the government and industry. As van Wolfern writes, "...it is impossible in Japan to separate the state from the socioeconomic system." And while it is true that the state is somewhat involved in the socioeconomic system in most nations, the involvement in Japan occurs to a degree that exceeds the involvement of any other nation, except for the centrally controlled economies. Unlimited industrial expansion is the consensus industrial policy of this partnership, to the single-minded exclusion of all else. In the United States and most other Western nations, other objectives such as defense and social-agenda items compete with industrial expansion, but not in Japan. Roadblocks to industrial expansion, such as the export-control problem faced by American firms [2], simply do not exist in Japan. And government-supported projects specifically intended to foster industrial growth are a commonly used government method of supporting

industry, the best-known example being the Super-Speed Computer System being supported by the Government of Japan [28]. To cite the large American R&D investment in defense as an equivalent policy in the United States is a categorical error: this kind of investment is aimed at another objective, and any contribution to industrial expansion is minimal and incidental.

A third myth about Japan is that competition by Japanese firms is the same as competition by the firms of other nations. The difference is what Peter Drucker calls "adversarial" trade as contrasted to "competitive" trade. In competitive trade, a country typically exports the same type of products that it manufactures with the aim of getting a share of a market, as case in point being the reciprocal trade in automobiles between Germany and the U.S. In adversarial trade, the objective is not just a share of the market but the market itself--a case in point being the continuing destruction by Japan of the American and European semiconductor industries. Japanese firms were competing well but not overwhelmingly with the firms of other nations through the mid-1970s. In 1976, MITI (the Ministry of International Trade and Industry) sponsored a VLSI Project that included Japan's six largest semiconductor manufacturers, with the goal of creating the technology for a one megabit chip and thereby giving Japanese firms a dominant share of the market. At that time, the standard was the 16K chip, but through the structural advantage gained from this government-industrial cooperative project Japan now leads the world in this field, with the impending demise of many companies in Europe and the USA. These other nations are now attempting to use cooperative research to regain lost ground, but they fail to understand the scope of the problem, which includes not only an adversarial national industrial policy but also adversarial corporate practices, including "dumping" and "anticipatory pricing," as noted in Section 3.4.

A final example of structural problems is to be found in Japan's "free ride" in defense and foreign aid. Japan's well known limit of 1 percent of its GNP for defense contrasts with some 6 to 7 percent for the U.S. and typically 3 to 5 percent in Europe [11]. Added to this is the fact that Japan does not carry its full share of the foreign-aid burden: whereas the U.S. spends $800 per capita on defense and foreign aid, Japan spends only $135. It has been suggested that Japan could begin to carry its foreign aid burden through an Asian version of the Marshall Plan [21], but this would be contradictory to the single Japanese goal of industrial expansion, and thus Japan has not picked up on this idea. These "savings" for Japan effectively lower the tax burden of Japanese firms which can then invest these funds in new product developments.

5.2 Government relations.

Specific implications of a partnership economy in which the government has as its primary objective the expansion of industry can be seen in (1) the differing effects of export controls in the two nations, (2) performance of government-supported supercomputer research R&D in Japan but not in the USA, and (3) the closure of the government market in Japan to US firms.

First, Cray Research reports that export controls in the U.S. are costing about 145 days for approval on the average, whereas in Japan the same approval takes less than 30 days. This fact is being used by Japanese firms in marketing their supercomputers: customers are being told by the Japanese that these delays are an unavoidable part of the American offerings but not of the Japanese offerings, and that there inordinate delays are possible, such as the 300-day delay suffered by the University of Stuttgart in Germany in obtaining approval for its Cray computer.

Second, in Japan, vendors of supercomputers are doing research for Government-sponsored supercomputer projects that are set up to benefit the industrial firms, but this is not true in the USA. This has the benefit of providing direct and specific government subsidies for supercomputer research in Japan, whereas the American companies must bear this burden out of thier own resources. In other words, the American firms are competing against not just the giant Japanese firms themselves but against the combination of the Government of Japan *and* industrial firms. Some specific cases in point are the Super-Speed Computer Project, supported by $100 million from MITI, and the Next-Generation Industries Project to develop advanced components needed by Japanese supercomputer firms, among other objectives.

Finally, the de facto [= true in the real world, whether it is true on paper or not] closure of Japan's government market (including their universities) to American supercomputer firms is in contrast to the openness of the American government market for the Japanese. Of the few (seven) American supercomputers installed in Japan, *none* is installed at a government site. Several Japanese universities have expressed interest in acquisition of a Cray supercomputer, but during the procurement process they have been advised that this would cause "political" problems for the funding of their universities if they should actually acquire an American supercomputer, and they have uniformly retreated from such acquisitions [10]. [NOTE: The Japanese will attempt to counter this argument by pointing to the installation of the IBM 3090 high-end mainframe as evidence of the openness of their supercomputer market, but this computer is not considered part of the supercomputer market by either IBM or American supercomputer firms.]

5.3 Cultural driving forces.

We have previously pointed out [see reference 8] the effects of the Japanese cultural driving forces in the market place, including their intense work and education ethic, their management style, and their three sacred treasures (lifetime employment, *nenko* reward system [= age priorities], and enterprise unionism).

6. SUMMARY AND CONCLUSIONS

6.1 Current market status.

- By its large market share, its timing of new product introductions, its strong marketing infrastructure, and its technical leadership, Cray Research has been able to hold off the threat to the supercomputer market by the Japanese competitors.
- As ETA Systems brings its ETA^{10} systems to market, their market share will be better protected against incursion by Japanese vendors. However, this is a narrow window in time that could be closed by either new products from Cray Research or the Japanese firms, and the delays being experienced by ETA in bringing this system to market are a serious threat to their very survival.
- Japanese vendors have largely but not completely captured the Japanese market through the traditional "buy Japanese" national bias of the Japanese culture and pressures from the Government of Japan. The Japanese government market is closed to American supercomputer firms. The American vendors been able to place their products only in private Japanese firms.

6.2 Key issues for the future. The main threats to the American supercomputer vendors include the following.

- The vertical integration of the Japanese supercomputer vendors that gives them control over the development of high-performance components and therefore a timing advantage in introducing new generations of supercomputers; the American firms must solve the problem of access to high-performance components in a timely manner.
- The large Japanese firms attempt to "buy" this market away from the small American firms by "anticipatory pricing"; it is difficult to see how to prevent this other than through action by the American government.
- The export control problems being faced by American firms but not by the Japanese firms; this problem must be solved by the American government.
- The closure of the government market in Japan to U.S. supercomputer firms; the American government should assure that American supercomputer vendors find a market in Japan that is as open as is the American supercomputer market.

7. REFERENCES AND BIBLIOGRAPHY

1. Karel G. van Wolferen, "The Japan Problem," *Foreign Affairs*, Winter 1986/87, pp. 288-303.

2. Colin Norman, "Academy Panel Blasts U.S. Export Controls," *Science*, Vol. 235, 23 January 1987, pp. 424-425.

3. U.S. Congress, Office of Technology Assessment, *Supercomputers: Government Plans & Policies--A Background Paper*, OTA-BP-CIT-31 (Washington, DC: U.S. Government Print Office, March 1986).

4. Stuart Auerbach, "U.S. Ponders Computer Trade Charge," *Washington Post*, February 5, 1986, pp. F1, F12.

5. Houston Area Research Center (HARC) News Release, January 15, 1986.

6. John W. Wilson, et al., "Is It Too Late To Save The U.S. Semiconductor Industry?" *Business Week*, August 18, 1986, pp. 62-67.

7. P. D. Lax, "Report of the Panel on Large-Scale Computing in Science and Engineering," National Science Foundation report NSF 82-13 (1983).

8. W. J. Worlton, "Japanese Supercomputer Initiatives," *Frontiers of Supercomputing*, N. Metropolis, D. H. Sharp, W. J. Worlton, K. R. Ames (ed's), (University of California Press: Berkeley, 1986).

9. B. Buzbee, R. H. Ewald, and W. J. Worlton, "Japanese Supercomputer Technology," *Science*, Vol. 218, 17 December 1982, pp. 1189-1193.

10. Private communication, Suzanne P. Tichenor (Cray Research, Inc.) to Michael B. Smith (Deputy U.S. Trade Representative), January 9, 1987.

11. S. J. Solarz, "A Search for Balance," *Foreign Policy*, 49 (Winter 1982-83), pp. 75-92.

12. Lewis J. Lord, et al., "The Brain Battle," *U.S. News and World Report*, January 19, 1987, pp. 58-64.

13. Kenneth Fleet, "Treasurey joins attack on Japanese barriers," *The Times*, February 12, 1987, p. 21.

14. Kimio Ohno, "Japanese Supercomputers and Molecular Orbital Calculations," Supercomputer Simulations in Chemistry, M. Dupuis, Ed., (Springer-Verlag: Berlin, 1986), pp. 49-54.

15. "Japanese Electronics: NEC," *The Economist*, April 12, 1986, pp. 73-76.

16. Willie Shatz and Tom Murtha, "Supercomputers: Planting the Flag," *Datamation*, May 1, 1986, pp. 24-28.

17. David E. Sanger, "Japan's Fast-Computer Plans Falter," *International Herald Tribune*, November 29-30, 1986, pp. 1,13.

18. James A. Martin, "Chip merger concerns U.S.," *Computerworld*, November 10, 1986, p. 126.

19. "Special Report: Japan's technology agenda," *High Technology*, August 1985, pp. 21-67.

20. Peter Drucker, "Doing Business in Japan Isn't 'Just Done'," *The Wall Street Journal*, July 18, 1985.

21. Sam Nakagama, "Let Japan Put Its Money Where Its Might Was," *The Wall Street Journal*, October 18, 1985, p. 30.

22. John Adam, "High technology called key to raising U.S. productivity," *The IEEE Institute*, November 1985, p. 6.

23. W. David Gardner, "Japanese Prepare New Wave of Assaults on U.S. Market," *Information Week*, September 2, 1985, pp. 30-37.

24. Mark L. Goldstein, "Yamamoto's Rising Sun," *Industry Week*, April 15, 1985, pp. 63-64.

25. Sidney Fernbach, "Supercomputer research and development in the U.S. and Japan: an update," *The IEEE Institute*, April 1985, p. 5.

26. Sachio Kamiya, et al., "Practical Vectorization Techniques for the 'FACOM VP'," *Information Processing 83*, R. E. A. Mason (ed.), (Elsevier Science Publishers (North Holland), 1983, pp. 389-394.

27. Charles L. Cohen and George Leopold, "NEC's Latest CPU Keeps Honeywell Line Going," *Electronics*, February 17, 1986, pp. 15-16.

28. Hiroshi Kashiwagi, "The Japanese Super-Speed Computer Project," *Future Generation Computer Systems*, Vol. 1, No. 3, February 1985, pp. 153-160.

29. Hisashi Horikoshi and Rasuhiro Inagami, "Japanese Supercomputers: Overview and Perspective," *Proc. IEEE Int. Conf. on Computer Design: VLSI in Computers (ICCD '84)*, pp. 227-231.

30. Tim Carrington and Robert Greenberger, "Fight Over India's Bid For Computer Shows Disarray of U.S. Policy," *The Wall Street Journal*, February 24, 1987, pp. 1, 24.

31. James A. Martin, "Japan ignoring chip trade accord," *Computerworld*, February 23, 1987, pp. 83, 87.

Emerging Supercomputer Architectures

Paul C. Messina
California Institute of Technology°

I. Introduction

This paper will examine the current and near future trends for commercially available high-performance computers with architectures that differ from the mainstream "supercomputer" systems in use for the last few years. These emerging supercomputer architectures are just beginning to have an impact on the field of high performance computing.

For some time it has been noted that sequential machines are approaching fundamental limits in speed imposed by the speed of light and heat transfer. This observation is coupled with the statement that significantly higher performance can only be achieved by decomposing a program into multiple parts and executing them concurrently on multiple hardware. Today's supercomputers achieve their performance through replication of certain components so that regular operations on vectors of operands can be performed at much higher speeds than operations on single pairs of operands. The well-known vector machines are of this type. Several of the vector computers also feature multiple CPUs, at present up to four, each with vector capability as well as the ability to access a large shared memory. In the context of this paper, these machines have "traditional" supercomputer architectures. Supercomputers manufactured in the United States include various Cray Research computers, Control Data Cyber 205s, and the ETA-10. Several Japanese vendors (Fujitsu, Hitachi, and NEC) also

° Part of the preparation of this paper was performed at Argonne National Laboratory, the author's former employer.

manufacture vector supercomputers. Amdahl Corporation, a U.S. company, markets the Fujitsu line of vector supercomputers and NEC markets its vector computers in the United States. At present, Hitachi vector supercomputers are not marketed in the U.S.

Over the last few years, computers whose architectures feature a greater level of parallelism than vector supercomputers have been introduced by commercial vendors. Some have modest performance at this time but promise to increase dramatically in capability. Others already have potential performance that would put them in the supercomputer category. It is these systems that are the topic of this paper. While none of them is being used as a general-purpose supercomputer today, there are signs that within a year or two that will change. This paper will focus on commercially produced computers. A few computers that are under development as research projects will be mentioned where it seems appropriate, but most architectures that are still at the basic research stage will not be discussed here.

II. Characteristics of Today's Supercomputers

To set the stage for examining the emerging supercomputer systems, we first survey briefly the present crop of supercomputers. In this paper, the term supercomputer will be used to mean the currently fastest general-purpose scientific computer. Performance is usually defined in terms of arithmetic operations on 64-bit operands. In practice, several systems will be called supercomputers at any point in time. It is not possible to say unequivocally that one machine is faster than all others. This is due to the difficulty of measuring performance for these machines and the inescapable fact that even the "general-purpose" systems will perform much better on some problems than on others.

Supercomputer Software Environment

In addition to possessing the highest peak and achievable performance, today's supercomputers are expected to have operating systems and software environments that are sufficiently mature to allow productive use with moderate effort on the part of the user. The operating systems and related peripheral hardware must have high performance I/O capability to match the internal speeds of the machine. The OS permits both interactive and batch use and coordinates the simultaneous access of the system by hundreds of users. Software and hardware for high-speed network connections are available to connect the supercomputer with a variety of front-end systems and computer networks. Optimizing compilers for popular languages such as Fortran and occasionally C are available. The Fortran compiler is expected to perform a substantial amount of vectorization automatically. Libraries of graphics and mathematical routines are available, as are many application programs such as structural analysis and circuit design packages.

Before the emerging supercomputers can be considered seriously as viable alternatives to the traditional variety, their software environment must offer the above capabilities. Mere peak performance is not sufficient to qualify a system as a supercomputer, especially if its architecture is novel.

Supercomputer Performance Levels

What then is the level of performance that must be achieved to merit the designation of supercomputer? Peak speeds of systems like the Cray 2 approach two billion floating point operations per second (abbreviated Gigaflops, or more commonly Gflops). A few user programs have achieved over one Gflop. Perhaps the first well-known instance of this was accomplished on a Cray 2 using all four processors simultaneously on one program [1]. Nearly that level of performance has been obtained on an Amdahl 1400 with four vector units and a

NEC SX-2. The Amdahl 1400 achieved 940 Mflops on a two-dimensional filtering code used in seismic applications. This is noteworthy because the code performs a considerable amount of I/O to disk (18.75 MB in and out) and the timing runs were done while the system was running other production jobs. The same machine ran order 1000 64-bit matrix multiply at 1046 Mflops. Few user programs approach those speeds, but performance of 50 million floating point operations per second (Mflops) per CPU is fairly common. In the next three years traditional supercomputers are expected to have basic cycle times that are one half to one quarter of today's fastest systems and the number of processors in one system will grow from four to sixteen or perhaps as many as sixty-four.

With those estimates in mind, we can say that peak rates of at least one Gflop and achieved performance on real user programs of several hundred Mflops are needed to qualify for the supercomputer designation today. Three to five years from now we should see peak speeds in the 10—30 Gflop range and achievable performance of several Gflops for a wide variety of scientific and engineering computations. These then are the performance levels to keep in mind when assessing the emerging supercomputer architectures against the traditional ones.

Performance Measurement is Difficult

A factor that complicates the measurement, and therefore the definition, of supercomputer performance is that the ratio of peak to achieved performance is growing larger. On a vector machine, a ratio of 10:1 might be seen. In highly parallel architectures the ratio between peak and minimum speeds achieved can be even more dramatic. On a parallel system with 1000 processors, each of which has vector hardware that is 10 times faster than the scalar units, one could see a ratio of as much as 10,000:1 between peak performance and a worst case of single-thread code that did not vectorize or parallelize. To make matters worse, relatively small changes in the algorithm, its implementation, the source code, and the compiler

can result in large changes in performance. When poor performance is experienced, it is difficult to determine whether it is due to inherent design flaws in the hardware or a minor aspect of the software that could easily be modified. Even when implementation inefficiencies are ruled out, it is foolhardy to assume that the system under study is not well suited for certain computations. Someone might invent a different algorithm for the same computation that uses the hardware effectively.

New Definition of Supercomputer Performance Needed?

As suggested by the following analysis, the emerging architectures may, perhaps surprisingly, play a significant role in the supercomputer arena even if their performance is an order of magnitude slower than the traditional supercomputer systems. Most traditional supercomputers are in centers with hundreds of users. A single user can expect to get only a small percentage of the available time. It is unusual for one group to get as much as 10% of the time on a supercomputer. If the cost of a computer is low enough that a small research group can buy one for its exclusive use and the performance of that system is 10% that of the "real" supercomputer, the same class of computations can be undertaken on the slower system, assuming that its memory and peripheral devices are adequate. It is indeed the case that some of the new systems offer sufficiently low cost and high performance to satisfy the scenario above. For example, at Argonne National Laboratory, Don Sinclair achieves for QCD computations two-thirds the performance of a Cray X-MP single processor on a Star Technologies ST-100 array processor. Both the Cray X-MP and the ST-100 programs are highly tuned. Since Sinclair has exclusive use of the ST-100, he can perform on that system computations of the same or larger magnitude as on a Cray X-MP. One series of computations [2] has consumed over 4500 hours of ST-100 time, which for this program is equivalent to 3,000 Cray X-MP hours. The elapsed time to do the computations is comparable or shorter than for using the Cray, given that it is extremely unlikely that a user can get exclusive access to a Cray X-MP for

months at a time.

Therefore, some of the emerging high-performance systems will in effect change the operational definition of supercomputer performance levels. We may well need to start thinking of *flops available per year to a single user or group* as the measure of supercomputer capability.

III. What are the Emerging Architectures?

Computer architectures have evolved in several directions in the never-ending quest for speed and price-performance. Most of the architectural innovations are based on the strategy of concurrency. If several tasks can be performed at the same time, in parallel, then the overall problem can be solved in less time than if only one task is active at any point in time, given components of the same speed.

A Look Backwards

To put the emerging architectures into perspective, it will be helpful to examine briefly the history of architectural innovations aimed at high performance. In the 1960's high performance systems began to feature concurrency. Processing of instructions by the central processing unit was decomposed so that one instruction could be fetched from memory while another instruction was decoded and a third was executed. Floating point arithmetic operations are complicated and therefore are often much slower than other operations. For scientific computation, floating point operations often dominate the execution time, so there is considerable incentive to speed them up. In addition to instruction pre-fetching and decoding, floating point function units were decomposed into segments, each of which did part of the

arithmetic operation. These segmented functional units operate in pipeline fashion. Once an operand has entered the functional unit and has been operated on by the first segment, a second operand can enter and begin processing in the first segment while the first operand is in the second segment of the functional unit, and so on. In addition, multiple arithmetic functional units began to emerge, so that, for example, two floating point additions/subtractions and one floating point multiplication could all occur at once.

Access to memory and input/output operations received similar attention. Memory units were split into "banks" so that successive data accesses, each one to a separate bank, could be carried out with considerable overlapping. Cache memories were added to exploit the locality of instruction and data access that many programs exhibit. These memories are much smaller but substantially faster than main memories. If program instructions and the data they require are in cache memory, they can be fed to the instruction processing unit much faster than if they must be fetched from main memory. Specialized I/O processors were introduced so that much of the effort associated with I/O operations could be performed by the I/O processor, leaving the central processor free to execute other program instructions.

The 1960's also saw the birth of several new types of architectures. These architectures used a different approach to gain performance. Rather than design a single processor and memory system with many elaborate features for concurrency, multiple processors were combined to form a single system. The most straightforward of these approaches was to have two or four entire CPUs share memory and peripherals. Each processor ran programs independently of the others, so there was no gain in speed for any one program, but a larger number of programs could be executed in a given period of time. That is, greater throughput was achieved. Large mainframes built by IBM and CDC were among the first to offer this type of configuration. In the 1980's this arrangement has propagated to minis and superminicomputers, for example many of DEC's 8000 series models.

Parallelism aimed at increasing the speed with which a single program could be processed became a topic of attention in the late 1960's and early 1970's. The ILLIAC IV

employed 64 rather simple processors and connected them in such a way that they could work simultaneously on computations for the same program. In the ILLIAC IV, all processors operated the same instruction in lock step. Each processor had its own modest sized memory and could send messages to four neighboring processors. The ICL Distributed Array Processor (DAP) utilized a much larger number (4096) of processors, but each processor was extremely simple and had a small memory. Again one instruction at a time was issued, but all 4096 processors could execute (or ignore) it, each on its own data. (In the early 1980's the Goodyear Massively Parallel Processor went even further in this direction: it had 16,384 processors. We describe the DAP and MPP in more detail later in the paper.)

In the 1970's the trend of ever more elaborate single-processor systems continued. Vector machines such as the CDC Star, Cyber 203 and 205, and the Cray-1 employed the techniques for speeding up single processors mentioned earlier, but in addition had one or more pipelines that optimized the execution of regular operations on vectors of numbers. The Cyber 205 can have up to four such pipelines, so if several vectors can be operated on independently of each other, those operations can be performed simultaneously. These vector machines have become the dominant architecture for very high performance numerical computing. Today's supercomputers have the same basic architecture as the vector machines of the 1970's. Higher performance for a single vector system has been achieved through shorter basic cycle times and removing some bottlenecks. For example, the Cray X-MP has three paths to memory that can be active simultaneously rather than one path, as in the Cray 1s.

The Denelcor Heterogeneous Element Processor (HEP) was a sophisticated pipelined MIMD system developed under funding from the Army Ballistics Research Laboratory and marketed as a commercial product in the early 1980s. The HEP consisted of one to sixteen Program Execution Modules (PEMs) that were connected to globally shared memory via a pipelined switch. Each PEM had local memory and a large register set. An eight-stage pipeline executed all instructions (except floating point divide) in eight clock cycles. The basic cycle time was 100 nanoseconds, so once the pipeline was full a result was produced every

100 nanoseconds. Process creation and management (up to 64 per PEM) was done mostly in hardware. This design made possible the exploitation of very fine grain parallelism within a PEM, as well as larger grain decomposition among PEMs. The HEP was aimed at the supercomputer market but its initial implementation was not cost effective and software was immature. When Denelcor Corporation went bankrupt, there were four or five customer sites. Only one, Messerschmidt, used it for production work: controlling a flight simulator.

The supercomputers of the mid 1980's are vector machines. Systems from Cray and ETA feature two, four, or eight complete vector processors, all capable of accessing a large memory. In some cases there is a sizable memory that is local to each processor as well. The Japanese supercomputers are all single processor models but have optional multiple vector pipelines.

The New Architectures

It is difficult to devise a systematic taxonomy of computers with advanced architectures. Among the possible distinguishing features are:

- the mechanism for control of execution—program, dataflow, or demand-driven;
- sequential or parallel execution;
- single or multiple instruction streams (SIMD versus MIMD);
- homogeneous or heterogeneous processors;
- grain-size—the size of the units of work that can be performed in parallel;
- the method and topology of connecting processors to memory and processors to processors;

- the characteristics of each processor—e.g., bit-serial, microprocessor, long instruction word, scalar, vector, or specialized;

The major architectural trends in commercially available machines are (1) program-driven multiprocessor systems featuring bus-oriented connections between processors and memory, (2) multiprocessor systems with hypercube connection schemes, (3) wide-word machines, and (4) lattice connected systems. Many systems combine several of these architectural features. In many cases vector processors are available. Soon several of the shared memory systems will have variants in which groups of processors that share memory are linked to each other and perhaps to a global shared memory. These designs are sometimes called "cluster architectures." Today's commercial multiprocessor systems are homogeneous. The processors used are generally not designed especially for use in parallel architectures; rather, they are processors that were developed independently for general-purpose use.

Many advanced architecture commercial systems are not initially designed or marketed to compete in the supercomputer arena. However, the designs can generally be extended and developed to the point that supercomputer performance could be achieved at least at the hardware level. Developing the rest of the characteristics of a true supercomputer is more problematic. For example, considerable time and money is necessary to develop the requisite software base and high performance I/O capability. It appears to be mostly a question of corporate strategy as to whether the systems with innovative architectures try to compete with the traditional supercomputers.

Bus-connected Multiprocessors

Bus-connected multiprocessor systems have been introduced by several commercial vendors. Most of these systems use one or more high-speed buses to connect relatively inexpensive processors of slow or moderate speed to a memory that is directly addressable (shared) by all the processors. The first generation of several of these systems uses NS 32032 or MC

68020 processors. Encore's Multimax-1, Sequent Balance 21000, and Flexible Corporation's Flex/32 are examples of this type of system. Current models can house up to 30 processors, so their present processing power is on the order of a few Mflops. New models are being introduced this year that use faster microprocessors and other enhancements such as larger caches and faster buses. For example, the new Sequent Symmetry Series uses the Intel 80386, which is three to four times faster than the NS 32032 used previously. Even faster floating point performance will be available through the addition of a floating point option based on Weitek chips. With this option, each processor should achieve about 0.3 — 2.0 Mflops; with perfect speed-up, a 30-processor system might deliver 20 Mflops on real applications. At that level, while still not a high performance numerical computation machine, it would be twenty times faster than a conventional supermini like a VAX 8800.

In contrast, Alliant has a higher performance system, the FX/8, that includes up to eight custom vector processors and up to thirteen MC68020 microprocessors. The 68020 processors are used for non-computational tasks such as compilation, editing, and the operating system. The peak speed of the FX/8 with 8 vector processors is about 46 Mflops for 64-bit data. Typical performance is much lower, say about 10 Mflops, but that is impressive in that it rivals the supercomputers of a decade ago, such as the CDC 7600.

Systems of this type have proved to be cost effective alternatives to traditional superminis and minisupers, but are not in the supercomputer performance range. Bus connections cannot support very large numbers of processors, since the bus bandwidth must be shared by all the processors on the bus. Bus speeds become one limiting factor. However, the "clustering" approach can in principle yield systems with peak speeds in the supercomputer range. The Flex/32 is scalable to high performance configurations, in the sense that up to 1024 cabinets, each with 20 processors, can be linked together. In addition, there are research projects underway to link together other systems so that the resulting system will approach supercomputer speeds. The Cedar project at the University of Illinois at Urbana, is developing a system that will connect up to 64 Alliant FX/8s. This configuration would have a peak speed

of about 3 Gflops. The ULTRAMAX project, carried out jointly by Encore Corporation and Carnegie-Mellon University, will also use the strategy of connecting multiple clusters to achieve performance in the range of 10^9 instructions per second (Gips). The clusters will consist of Encore Corporation's Multimaxes, each with 20 processors. It has not yet been demonstrated that one can get high performance for real applications on cluster systems, but experience with distributed memory machines suggests that it will be possible.

Hypercube Architecture

The second popular advanced architecture features the hypercube connection scheme. With this architecture, each processor has private memory and is connected to other memory/processor combinations (known as "nodes" in the parlance of message passing architectures) via a high speed channel. In some variations of this architecture, subsets of nodes share memory. This architecture scales well to large numbers of processors, because in a hypercube with N nodes each node is connected to only $\log_2 N$ other nodes, where N must be a power of 2 ($N = 2^m$ for some m). As with bus-connected systems, most of the commercial offerings use rather slow microprocessors at each node. In early versions of these systems, the communication speed among processors has been so slow that it has further throttled the performance achieved.

Intel markets the iPSC line of hypercube architecture systems. There are three types of configurations. The standard system has up to 128 nodes; at each node is an Intel 80286 microprocessor with the Intel 80287 floating point coprocessor and 0.5 Mbytes of memory. Each node is capable of about 0.03 Mflops. The large memory version has the same processors but with 4.5 Mbytes of memory. Its maximum configuration is 64 nodes. Finally, there is a version with vector processors at each node. This model is offered in configurations with up to 64 nodes each with 1.5 Mbytes of memory. Architecturally it is possible to build a vector system with 128 nodes, but that configuration is not a standard product. The peak speed of the 64-node configuration is 424 Mflops (64-bit), fast enough to approach supercomputer

performance.

On a seismic modeling application, performance of 227 Mflops was achieved on a 32-node system with vector processors. The inner loop of the computation was microcoded by hand to achieve this performance. This computation was done with 32-bit accuracy.

The largest Intel vector system built to date has thirty-two nodes. A new model with faster node processors (Intel 80386) and more memory has been announced; each node will be about three times faster and new communications hardware and software should substantially increase the achievable performance. While the vector option will use the same board as at present and therefore will have the same peak speed, higher speeds should be possible on real application programs because of better communications and compilers. Fortran, C, and LISP are available.

A hypercube with up to 1024 processors, each approximately the speed of a VAX 11/750, is available from NCUBE Corporation. Each node has either 128K or 512K bytes of memory. The processors are a custom design and extremely compact. Sixty-four processors fit on a single board. Even though there is no special hardware for floating point arithmetic or vector processing, the largest configuration has peak speeds of over a billion instructions per second and two hundred Mflops. It also has potential for high-speed I/O: eight I/O channels can transmit data at 90 Mbytes/second bidirectionally. As was noted earlier, an important characteristic of supercomputers has been the ability to do high speed I/O. The NCUBE system achieves supercomputer performance on some computations. For a Monte Carlo photon transport program, a 64-node NCUBE ran at about the same speed as a Cray X-MP single processor for small problems and 1/12th the speed on large problems. The Cray code was vectorized. [3] Fortran and C are available.

Floating Point Systems has introduced the T-Series, in which modules, each containing eight array processors, are connected to each other with the hypercube scheme. The T-Series can be configured with up to 2^{14} processors. Each processor has a peak speed of 16 Mflops. The T-Series thus is potentially the most powerful computer system available: peak

performance of the maximum configuration is 262 Gflops. At present the only programming language available is Occam and the performance actually achieved is usually far short of the peak performance, due to its very slow scalar speed. However, on the biggest existing T-Series configuration, a 128-node system, hand-coded matrix multiplication routines have run at a speed of 1.2 gigaflops for matrices of order 1024. Two-dimensional convolution achieved similar speeds. Fortran and C will become available in the near future.

Ametek Corporation also has announced systems with the hypercube architecture. Its initial model used Intel 80286/80287 processors; maximum configuration was 256 nodes.

The hypercube connection scheme is also used by a radically different system: the Connection Machine, built by Thinking Machines Corporation. The Connection Machine is massively parallel: its maximum configuration has 65,536 processors. They are quite different from the other processors discussed to date: each is a one-bit computer. In the original model, the CM-1, each processor has only 512 bytes of memory and floating point operations are performed by groups of processors (microcode is provided to do this). The Connection Machine Model 2 (CM-2) has 65,536 bit-serial nodes and 2048 floating-point units (based on a Weitek chip set). In the CM-2, each processor has 8K bytes of memory. The system is built from chips that contain 16 bit-serial processors each. Within a chip all 16 processors are connected. Connections between the chips are in the hypercube topology. Unlike the other hypercubes, all of which are MIMD, the Connection Machine is an SIMD machine. All processors receive the same instruction each cycle, which they may ignore depending on the setting of a flag bit. The Connection Machine architecture is therefore similar to that of the DAP and MPP described below. They also use bit-serial processors and have the SIMD control model. A key difference is that the MPP and DAP use lattice connections whereas the Connection Machine uses the hypercube topology. The only languages currently available on the Connection Machine are special versions of LISP and C, but Fortran will be added in a few months. While the original Connection Machine was designed with artificial intelligence applications in mind, it can perform floating point arithmetic at high rates. Speeds of 50

Mflops have been achieved on a 32,768 node configuration of the first model and it appears that 100 Mflops would be reached with a 65,536 node system. The CM-2 is claimed to achieve 2.5 Gflops for 64-bit matrix multiply and 5 Gflops for 64-bit dot product. This machine has a sophisticated programming environment but it is oriented to programming in an enhanced version of LISP, a language that is heavily used in artificial intelligence work but is virtually unknown in numerical programming. An enhanced version of C is also available.

Finally, although not a commercial product, the Mark IIIfp hypercube system (that Jet Propulsion Laboratory is building as a research project) is worth mentioning. The Mark IIIfp can be configured with up to 256 nodes. Each node has two MC 68020 microprocessors, one for computation and one for communications; four megabytes of memory; and a pipelined floating point unit with relatively high performance, currently estimated at 10 Mflops for 64-bit arithmetic. The floating-point board uses the new Weitek XL series of chips. Preliminary timings on the 32-bit version of the chips on the Mark IIIfp board (the 64-bit version of the chips is not yet available) yielded speeds as high as 16 Mflops for a hand-coded assembler code to multiply two 3×3 complex matrices. This computational kernel is at the heart of many QCD computations. The 128-node configuration under construction will therefore have a peak speed of over one Gflop, putting it in the supercomputer performance level.

Wide Instruction Word Machines

It was noted in an earlier section that one technique for gaining higher performance in computers was to have more than one functional unit in the CPU and to find ways to utilize them concurrently. This approach takes advantage of parallelism at a very fine level, that of individual arithmetic or logical operations. Array processors of the 1970's exploited this approach, the Floating Point Systems product line being the most successful. On the FPS 164 and 264 attached processors, array indexing, loop counting, and data fetching from memory can be performed simultaneously with arithmetic operations. Wide instruction words are used

to specify the simultaneous operations. Furthermore, up to 15 MAX (Matrix Algebra Accelerator) boards can be added to these machines. Each MAX board has two multipliers and two adders. Compiler technology did not yield code that took full advantage of the multiple units; therefore, intricate hand coding, sometimes in microcode, was necessary to get the high performance. However, when this could be done, the cost effectiveness of this type of hardware was impressive. Enrico Clementi of IBM-Kingston was able to perform large computations by running on multiple FPS 164s controlled by an IBM mainframe. This success led Cornell University to create a high-performance parallel system by attaching two FPS 164 and five 264 processors to an IBM 3090-400 computer.

Another successful wide instruction computer is the ST-100 array processor. With a 40 nanosecond clock cycle and four independent programmable processors, its peak performance is around 100 Mflops in single precision (32-bit) arithmetic. A separate processor is dedicated to each of the following functions: external data flow, internal data flow, arithmetic processing, and synchronization. Arithmetic processing is performed by 32-bit floating point arithmetic, pipelined functional units: two adders, two multipliers, and a 480 nanosecond divide/square root functional unit. Several memory references and logical operations and four arithmetic operations may be started in each machine cycle. Since it does not support 64-bit floating point arithmetic and at present has no compiler for a high level language, it cannot be considered a general purpose scientific computer. Nevertheless it provides evidence that this type of architecture can be cost effective; with careful programming it has achieved a sustained performance two-thirds that of a Cray X-MP processor for important scientific computations like QCD.

The latest commercial wide instruction word machines carry that architectural trend much farther. The CHoPP by Sullivan Computer will have a 256-bit "superinstruction" that is equivalent to nine instructions on a conventional system. Four functional units can perform address computations while four other functional units can perform full instructions, including floating point arithmetic. The ninth unit is reserved for branching. The CHoPP will be

available in configurations of four to sixteen CPUs with a large shared memory. Peak performance for the four CPU configuration is projected to be 270 Mflops; simulated performance for Livermore Loops 1 — 14 is 81 Mflops.

The Multiflow computer uses an instruction word that can be as wide as 1024 bits. Each instruction can start as many as twenty-eight primitive operations on reduced instruction set processors. Orchestration of the actions of that many low-level processors requires a highly sophisticated compiler. When it is successful, parallelism can be exploited at a very fine level. Other than the sophistication of the compiler technology, the software environment is conventional. Compilers for Fortran and C are provided. The UNIX operating system and TCP/IP networking protocols are used. Performance data are not yet generally available. At the time this paper was written, the first Multiflow computer was undergoing field tests. The smallest Multiflow system, the TRACE 7/200 which has a 256-bit instruction word, reached 6.0 Mflops on the full precision LINPACK benchmark. For a C program that does symbolic manipulations, but no floating-point arithmetic, it ran 16 times as fast as a VAX 11/780.

Lattice Connected Machines

SIMD machines consisting of large numbers of lattice-connected bit-serial processors constitute a high-performance advanced architecture that has been in use for some time yet can still be regarded as advanced. The Goodyear Massively Parallel Processor (MPP), the International Computers Limited (ICL) Distributed Array Processor (DAP), and the recently announced DAP-3 all fall in this category. They are similar to the Connection Machine but have a lower-dimensional connection network among processors.

The ICL DAP has been available since the late 1970's. Half a dozen machines have been installed. As was mentioned earlier, the DAP is an SIMD lockstep machine that operates on multiple data a bit at a time. Through programming, arithmetic can be carried out in variable precisions. The processing elements form a grid with nearest neighbor connections.

Typical configurations have a 64 by 64 grid of processors, each with 2048 bytes of memory. An operation can be performed on each processing element at each clock cycle of 200 nanoseconds. Masking enables or inhibits execution of the (same) instruction for each processor. Fortran is available through a cross-compiler that runs on the host ICL 2900. A 32 by 32 version of the DAP, designated the DAP-3, has been announced by Active Memory Technology Inc. of Atlanta, Georgia. Memory sizes range from 64K to one megabit per processor. As of this writing, none has been installed.

Goodyear Aerospace Corporation manufactures the Massively Parallel Processor (MPP). The only MPP built to date was delivered to NASA's Goddard Space Flight Center in May 1983. Its 16,384 processing elements are bit-serial processors arranged in a 128 by 128 grid with nearest-neighbor connections. Until the Connection Machine CM-2 was announced, the MPP was the highest performing SIMD system available. For image processing tasks it has performed over one billion operations per second. In numerical computations like solving partial differential equations, hundreds of Mflops have been achieved. The MPP utilizes Parallel Pascal as its high-level language.

The British company Inmos produces the Transputer, a 7.5 MIPS chip with four communications channels built-in, which make it well suited as a building block for lattice-connected systems. A new version with 1 Mflop performance will be available in September. Both Inmos and another British company, Meiko, market Transputer-based systems. It is expected that the University of Edinburgh will soon get a system with 1024 nodes of the new version. That configuration would have over 1 Gflop peak performance and 2 Gbytes of memory.

Other Architectures

BBN Advanced Computers Inc. produces the Butterfly family of computers. These systems consist of microprocessors and memory units that are connected to each other via a

specially-designed switch. Although all memory is local to the processor to which it is attached, each processor has access every other processor's memory through the switch. This general type of connection scheme is found in several research computers, such as the NYU Ultracomputer and IBM's RP3, but at present is unique to the Butterfly among commercial systems. Butterfly processor nodes consist of Motorola 68000 or 68020 microprocessors with one to four megabytes of memory. The largest configuration built to date has 256 processors. On that system typical numerical algorithms like matrix multiply have achieved speed-ups as high as 234, for an efficiency of 0.91. A future system known as the Monarch is under development as a DARPA-sponsored research project. This system could have as many as 8,192 processors. The first configuration is expected to have 1024 processors, each capable of one MIP and one Mflop. Fortran, C, and LISP are available.

There is at least one commercial computer that uses a dataflow architecture, the Loral Dataflo. Its maximum configuration is 256 nodes. At each node are two National Semiconductor NS32016 microprocessors, one for data management and one for program execution. The latter has an NS floating point co-processor. There is shared memory as well as memory local to each node. The system is programmed by supplying a data graph description plus a graph node program written in Fortran or C. The grain size for the system is said to be approximately the size of a procedure, say 60 to 100 lines of source code. Little is known about performance of this system; there may not be any customer sites yet. It is clear that the present generation of this system cannot have very high performance. The microprocessors used are quite slow; even the maximal 256-node configuration would have a peak speed of about 10 Mflops with 100% efficiency. It is included in this paper primarily because of its unusual architecture for a commercial computer.

The CYBERPLUS computer built by CDC is another unusual system. It is a multiple parallel processor system with a ring bus architecture. This computer features an 800 Megabits/second transfer rate with a read and a write possible between processors at this sustained rate. Two CYBERPLUS processor models are available: 16-bit integer and 64-bit

floating point. The floating point processor has a peak performance of 65 Mflops in 64-bit mode. Each processor has a 20 nanosecond cycle time. There are fifteen independent functional units and it is possible to start several operations in each instruction. Thus it has a wide instruction word architecture as well. Up to 16 CYBERPLUS processors can be connected to a ring. Within one ring all processors operate autonomously and may execute each clock cycle. Processor Memory Interface allows direct reading and writing of the memory of any processor by another processor on the ring every machine cycle. Up to 16 rings can be connected to a Cyber 800 series host computer. There are three distinct memory systems of different sizes and speeds. The host CDC computer (using the NOS 2 operating system) loads code into the processors, transmits data from host to processors, and starts and stops each processor's task. Software includes a cross assembler and an ANSI 77 Fortran cross-compiler. Notwithstanding the richness of the CYBERPLUS architecture, it has not established itself as a powerful general-purpose numerical computer. The peak performance of a large CYBERPLUS configuration is very high: 64 CYBERPLUS systems linked together are claimed to perform 16 Gflops (probably 32 bits) on a signal processing application. However, the existing customer sites have not been able to reach a substantial percentage of that performance for scientific computations.

IV. Software Environment

Languages and Compilers

There is a greater variety of languages and compiler technologies in the emerging supercomputers than in the traditional ones. Although most systems have (or plan to add) a Fortran compiler, examples of the diversity are the use of Occam in the FPS T-Series as the high-level language; special versions of LISP and C for the Connection Machine; and the existence of C and Pascal compilers in most systems.

The Fortran language for these systems is generally augmented by special syntax, system routine calls, or specially coded comments that provide additional information to the compiler.

In large-grain machines, compilers seldom aid in the parallel programming task. An exception is the Alliant's Fortran compiler that does automatic parallelization as well as automatic vectorization. Its ability to detect automatically vector operations and parallelism and to generate efficient code is respectable.

In contrast, compilers for fine-grain computers like the Multiflow and the Connection Machine perform a great deal of the parallelization automatically and are crucial to the performance achieved on those systems.

Operating System and Networking Trends

Virtually all the new high-performance computers (or their hosts) have some version of the UNIX operating system. In many cases the advanced system itself has a minimal operating system, but requires a host that runs UNIX. There are some exceptions to this pattern. The Connection Machine uses a Symbolics LISP machine as a host, though even here support for a VAX/UNIX host will soon become available. The DAP-3 can be hosted by a VAX running VMS, as well as a VAX or Sun running UNIX. The FPS array processors can be hosted by an IBM/MVS front-end or a VAX/VMS front-end, as well as UNIX systems. The FPS T-Series is hosted by a micro-VAX with the VMS operating system. The ST-100 can be hosted by VAX/VMS, micro-VAX, Gould, and Perkin-Elmer systems. The NCUBE system, since it requires no external host, runs most of the operating system on its Host Board with parts of the operating system running on the nodes. The NCUBE operating system, called AXIS, is similar to UNIX.

Since in most cases the advanced architecture machines have a host that runs a variant of UNIX or VMS, TCP/IP and DECNET networking protocols are generally available, with the former being the most common.

Application Software

There is limited availability of mathematical software, graphics libraries, and application software (such as engineering packages) for the advanced architecture computers. This is understandable since so few copies of any model have been installed for production use. For medium performance systems like Sequent, Encore, and Alliant, mathematical software libraries are under development. Sequent and Encore have begun to offer database packages. Engineering application packages are generally not available. This lack of application software is a significant obstacle to the use of the new systems in a supercomputer facility for a general user community. It is less important for basic research that requires supercomputing capability. For these uses, the programs are often written entirely by the research group responsible for the computation.

V. Current Use of Advanced Architecture High-Performance Computers

Computers with advanced architectures, as defined in this paper, are being used to perform large computations. A few advanced architecture systems have achieved performance on real user programs that is high enough to approach supercomputer levels but no highly parallel machine is in use as a supercomputer in the traditional sense. The Goodyear MPP is probably the closest to that status, but to achieve good performance on the MPP requires a substantial effort by the user and frequently the use of assembler language.

In most cases, the machines that are the topic of this paper do not exist in large enough configurations to be considered supercomputers. Furthermore, they were typically acquired for exploratory investigations on the use of parallel architectures. As a result, there are few systems that are used as workhorses for large-scale, production applications. On the other hand, the experience to date is encouraging for those who expect to obtain highly cost effective and high performance computing resources through advanced architectures.

It is interesting to note that the chief cause of the moderate performance of today's advanced architecture computers is the modest power of the processors used, rather than inability to use the systems with high efficiency. As was noted earlier, speed-ups of 234 have been measured on a 256-processor Butterfly. The SIMD fine-grain systems manage to achieve extremely high degrees of concurrency. Thus it appears possible that well-balanced, highly parallel architectures will indeed become supercomputers when they incorporate individual processors with higher performance than those currently used or, in the case of fine-grain SIMD machines, when hundreds of thousands or millions of processors are used.

The experience data presented here are necessarily sparse and vague; it is intended to give a feel for what has been achieved, rather than a systematic survey of all use of innovative systems. As was mentioned in the description of the NCUBE, William Martin at the University of Michigan has done Monte Carlo photon transport computations on a 64-node NCUBE at speeds that were a substantial fraction of a Cray X-MP-48. On a 16-node Intel iPSC-VX system, a system of 1000 dense linear equations was solved at a rate of 11 Mflops. On a seismic modeling application, performance of 227 Mflops (32 bits only) was achieved on a 32-node system with vector processors, but microcoding was used. With the new version of the iPSC and a 64-node configuration of the iPSC-VX, speeds of over 100 Mflops should be possible. An Intel iPSC 32-node system (but without the vector boards) is being used to process seismic data at the Christian Michelsen Institute in Bergen, Norway. The DAP's main applications are in lattice gauge theory and molecular dynamics. It is particularly powerful on the Ising model because of its bit arithmetic. It is also used in many Monte Carlo calculations and in image processing where the major problem is in data movement rather than processing speed. For some specialized applications, the DAP will outperform a Cray-1.

In brief, as was mentioned at the beginning of this section, this author is not aware of any supercomputer facilities that utilize systems with the novel architectures described in this paper. This situation may change in the near future as systems mature further, techniques for their use become more widely known, and a few institutions act as pioneers.

At least two institutions have plans to use highly parallel architecture computers for supercomputer-level work: Cornell and Caltech.

Cornell University has announced intentions of creating a massively parallel supercomputer facility. Indeed the IBM/FPS computer combination that is in place now is being used for large-scale scientific computing. Scientists from many disciplines have used the system to carry out major computations. The configuration consists of a conglomerate of commercial systems that provide a modest level of concurrent processing capability, rather than a massively parallel, unified architecture: the IBM 3090s provide four vector units and mainframes to control and host the various FPS array processors.

The California Institute of Technology has been using hypercubes for scientific computations since 1983. The Caltech-designed and built Cosmic Cube and Mark II systems have been used for QCD computations that have led to a number of physics publications, even though those systems are much slower than supercomputers: the 128-node Mark II runs at about 5 Mflops. This was possible because the machines could be dedicated to one computation for long periods. Scientists from other disciplines have also used the Caltech hypercubes for their work. In addition to the locally built systems, commercial hypercubes from Intel and NCUBE have been acquired and used by Caltech. In collaboration with the Jet Propulsion Laboratory (JPL) the Mark III hypercube has been built and a version with high-speed floating point, the Mark IIIfp, is under construction. When the new floating point version is completed, the 128-node Mark IIIfp will have a peak speed of over one Gflop. The NCUBE system will be expanded to 512 nodes the summer of 1987 and possibly 1024 nodes later. These systems form the hardware basis for the recently formed Concurrent Supercomputer Initiative at Caltech (CSIC), a project whose goal is to create a supercomputer facility based on highly concurrent architectures.

VI. Limits of Scalability

How big a computer can be built and used in the next three to five years? This question is asked increasingly often as large-scale computing pervades more and more disciplines and practitioners of those disciplines. Scientists and engineers always seem to need several orders of magnitude more computing power than is available. How likely is it that in the next three to five years it will be possible to provide an increment in computing power of two or three orders of magnitude? Earlier in this article current supercomputers were characterized as having a peak speed of a few Gflops and an attainable speed of several hundred Mflops. Memory sizes range from hundreds of megabytes to a couple of gigabytes. An increase in speed of two to three orders of magnitude would result in systems with a peak speed of 1,000 Gflops and attainable speeds of 100 Gflops or more. If memory sizes kept pace, these systems would have directly addressable memories of tens to hundreds of gigabytes. Can such systems be built in the near future? What type of architecture might they have? If they incorporate massive parallelism, is there any hope that they can be programmed effectively, so that single large scientific computations could use them efficiently? These are difficult questions to answer conclusively, but some rough estimates can be made with a simplified analysis of the construction and usability issues. We will restrict the analysis that follows to traditional scientific computations that are dominated by floating-point operations.

A widely accepted estimate of further speed increases for a single sequential processor is only one order of magnitude more than the fastest currently available. That estimate is based on advances in device and chip technology. Given that some systems currently have cycle times of as little as four nanoseconds, cycle times of a few hundred picoseconds would be possible. Computers built with such technology are surely more than three years away, since devices of that speed are not yet available as commercial products. We assume that in the near future device speeds for the fastest chips available will increase at most by a factor of four, and therefore will not be the key to getting two to three orders of magnitude speed increase.

We also assume that tying together 1,000 vector processors each with a peak speed of 1 Gflop is not a realistic way to create a system with a peak speed of 1,000 Gflops. Even if one used a distributed memory approach with a simple connection scheme such as ring topology, the cost would be prohibitive. Vector processors that run at 1 Gflop are expensive; the communications channels to link them would also be expensive. If the cost per vector processor were reduced to $1 million and communication channels were cheap, it is still unlikely that any institution would spend over $1,000 million for the system. The physical size of such a configuration would require a large building as well. This naive analysis suggests that in the near term one would have to use large numbers of much smaller and cheaper processors to build a very high performance system. For example, the new XL-series floating-point and integer arithmetic chip set is just becoming available from Weitek. At present only 32-bit versions are in production; early next year the 64-bit versions will be available. These chips have a peak speed of over twenty Mflops. As was mentioned in the description of the Mark IIIfp hypercube with Weitek chips, speeds of 16 Mflops have been measured on hand-coded assembler code to multiply two 3×3 complex matrices. Five Mflops ought to be possible on a reasonably wide range of computations. Other vendors may have or are developing chips with similar or better performance. In keeping with the style of this article, the Weitek chip set is used as an example and proof of existence, not as an indicator of the most advanced commercial product available.

How fast a system could one build today with such components? It would require 50,000 such units to reach a peak speed of 1,000 Gflops. Can systems that big be put together? There are indications that they can be, at least if a distributed memory architecture is used. The Connection Machine Model 2 has 65,536 bit-serial nodes and 2048 floating-point units (based on a different Weitek chip set). This system is not very big physically. Layouts for configurations with sixteen times as many processors have reportedly been worked out. Based on these considerations, it would appear possible to build a system with 32,768 floating-point units. If each has 20 Mflops peak performance, the system peak speed would be

655 Gflops. Based on early experience with the XL-series Weitek chips, well over 100 Gflops would be attainable for real applications on such a system. Using 5 Mflops as an estimate of achievable performance, a 32K-processor system would produce over 150 Gflops for highly parallelizable computations. If the price and size of this system were directly proportional to existing models, it would be realistic to build it.

The analysis above was based on just one possible system. Other components and connection schemes should also reach the 100 — 1,000 Gflop goals. Faster components are certain to be developed in the near future. In summary, it appears feasible to build systems with much higher performance.

Much larger memory sizes are also attainable, provided one can use relatively slow memory. In distributed memory architectures, memories need only be well matched in performance with the local processor (or processors, in the case of cluster architectures). Shared-memory architectures can also effectively use memory systems with components that are slow relative to the aggregate processor speed, provided a suitable pipeline or memory-access network is used. In shared-memory systems, once more than 4 Gbytes are used it will be necessary to devote more bits to addresses than at present. This is a cost, not feasibility, issue.

Having dealt with the hardware issues, admittedly in a most superficial manner, would it be possible to program systems with that much parallelism? Will the efficiency achieved be acceptable or will a variant of Amdahl's Law doom us to waste most of the hardware resources?

There is little doubt that it is possible to use effectively systems with a few hundred 32 or 64-bit processors or thousands of bit-serial processors. For typical scientific and engineering computations it has already been demonstrated that efficiencies of 70% or more can be achieved on systems of that type. The necessary algorithm redesign and re-implementation can often be done with moderate effort. High-level languages with some system subroutine calls are generally adequate; hand-coded assembler or microcode is not required.

Does the situation change for systems with thousands of 32 or 64-bit processors or millions of bit-serial processors? The answer depends on the computation to be carried out. For many scientific applications it appears feasible to use effectively arbitrarily large numbers of large-grain processors, provided the computation to be carried out is big enough. We will sketch out below the analysis and supporting evidence.

A number of real applications have been discovered to be "embarassingly parallel," that is, for sufficiently large problems, the computations can easily be decomposed to run on arbitrarily many processors with essentially no serial bottleneck. For such applications, the key question is whether it will be possible for each processor to get at the data it needs without incurring heavy overheads. On shared-memory systems this translates into whether one can design (and build at a reasonable cost) a processor-memory interconnection scheme that enables processors to get data from memory at a fast enough rate and without interference among themselves.

Shared-memory systems with thousands of processors have not yet been built. Research and development of effective interconnection networks is proceeding at several institutions but their near-term targets for the maximum configuration are at most 4096 processors. Even if massively parallel shared-memory systems are built, it is not known at this time how effectively the connection schemes, once built, will avoid bottlenecks.

For distributed-memory architectures, systems *have* been designed and in some cases built with many processors, so we assume that scalability of the hardware is not a key issue. Therefore, let us focus on a processor's ability to get the data it needs quickly enough. The communications channels among processors are typically slower than direct memory access but simpler to characterize, especially if there is no overlap of processing and communications. If the data are in the same node, i.e., in the processor's local memory, there should be no problem, assuming that the speed of the processor and memory at each node are well matched. With large memories at each node, the probability increases that the data are already in the node where they will be used. But there will be times when communication among nodes is

necessary, either to transfer data or to synchronize activities. How big is the communications overhead? We summarize here two analyses, one by E. Amdahl and one by G. Fox.

In a keynote address at the Supercomputing '87 Conference, May 1987, "Tempered Expectations in Massively Parallel Processing and Semiconductor Industry," E. Amdahl presented an analysis of the scalability of parallel systems. There was no paper published in the proceedings, so what follows is based on notes taken during his presentation. Amdahl used the hypercube architecture as the basis for his analysis because from a hardware standpoint it is the most promising for extending without limit the number of processors. Here is my summary of his analysis. Consider an N-node hypercube, $N = 2^n$. Let $W =$ the total (sequential) workload, let c be the communication and task switching load among processors, and let g be the "globality" of the data. The quantity g reflects how often it will be necessary for a node to get data from another node. The case $g = 1$, corresponds to random distribution of data, whereas if $g < 1$ the probability that data are "far away" from the node that needs them decreases exponentially as $g \to 0$. Assuming that the average number of hops per communication is

$$E_{AV} = \frac{g}{1+g} \log_2 N$$

and that one message per node is sent, then the workload on N nodes, including communications, is

$$W_N = W + \frac{cg}{1+g} N \log_2 N \qquad (1)$$

i.e., it is the original workload plus the communication overhead. It is assumed that communications is not overlapped with computation; this is the case in most of the current commercial distributed-memory systems. Let S be the speed of each processor. Then the execution time for one processor is

$$T_1 = W/S$$

and for N processors it is

$$T_N = \frac{W_N}{NS}$$

Restricting this analysis to the effects of communication, that is, assuming that the workload W is 100% parallelizable, then the relative performance (speed-up) on N nodes compared to one node is

$$R_N = \frac{T_1}{T_N}$$

or, substituting expression (1) for W_N and simplifying,

$$R_N = \frac{N}{1 + \frac{cg}{1+g}\frac{N}{W}\log_2 N} \tag{2}$$

and the efficiency is

$$\epsilon = \frac{1}{1 + \frac{cg}{1+g}\frac{N}{W}\log_2 N} \tag{3}$$

Amdahl used equation (2) to conclude that one can get arbitrarily high performance on hypercubes, provided the workload is large enough. (In his presentation, the number of processors was used as the variable to adjust to get higher speed-up. I prefer to use grain size and total workload as the variables; the results are the same.) What is needed is that $\frac{N}{W}$ grow more rapidly than than $\frac{N}{\log_2 N}$, say

$$\frac{W}{N} \approx N^{\delta-1}\frac{cg}{1+g}\log_2 N \ , \ \delta > 0 \tag{4}$$

Since $\frac{W}{N}$ is the workload per node, this is a measure of the grain size that must be used to get

speed-up. According to equation (4), no matter how large N is, one can get increased performance by adding more nodes, provided the problem is big enough that the grain size is suitable, that enough computation is done on each grain, and the communications load does not increase too rapidly with N.

Amdahl pointed out that although performance can be increased indefinitely, the efficiency may be low. For example, if $g = 1$, $N = 1024$, and the communications workload c is 0.1% as big as the computational workload W, then $\epsilon \approx 0.16$.

Amdahl also said that for certain computations efficiency could be kept high even for very large numbers of processors. He did not present the supporting analysis, but the following straightforward manipulations provide a basis for that statement. Clearly one wants the efficiency to be be as close to 1 as possible, even for very large N. That corresponds to

$$\frac{cg}{1+g}\log_2 N \frac{N}{W} \leq \delta$$

for some suitably small δ, $\delta > 0$. This will be true if

$$\frac{W}{N} \geq \delta^{-1}\left(\frac{cg}{1+g}\right)\log_2 N \qquad (5)$$

just as with speed-up, if the computational workload per node $\frac{W}{N}$ is large enough for the configuration used and the communications workload has a certain behavior. We need to characterize the type of workload for which (5) holds true. The left hand side of (5) can be made arbitrarily large by choosing a large enough workload W for a given value of N (or for a fixed workload choosing a suitably small N). Examine the right hand side of (5). δ is a constant and $\frac{g}{1+g}$ is always less than 1. The dominant behavior is that of $c\log_2 N$, essentially the communications workload per node. If c depends on N and W , and grows with the grain size $\frac{W}{N}$ sufficiently fast, then (5) will not hold and efficiency will be low. Otherwise, for sufficiently large grain size, efficiency will be high.

Fox [4, 5] had previously developed an analysis similar to that presented by Amdahl. The same qualitative results are obtained and empirical data are presented that support the conclusions. In addition, Fox noted that these results have implications on the memory size of each node and the amount of computation done on the data relative to the communication time required to transfer the data to the node where it will be used. Since this work has been published, only the key results will be be mentioned. In Fox's formulation, the following (machine-dependent) parameters are introduced

t_{calc} — the time required to perform a generic operation

t_{comm} — the time required to communicate one word from one node to another

For a given computation, the total calculation and communication per node are denoted

T_{calc} — the number of operations done in each node × t_{calc}

T_{comm} — the number of words transferred to/from each node × t_{comm}

The fractional communication overhead is defined as

$$f_C = \frac{T_{comm}}{T_{calc}}$$

and it is shown that the efficiency can be expressed as

$$\epsilon = \frac{1}{1 + f_C} \tag{6}$$

which is equation (3) with

$$f_C = \frac{cg}{1+g} \log_2 N \frac{N}{W} \tag{7}$$

This is not a superficial similarity, because $\frac{W}{N}$ is a measure of the calculation time per node, which is what T_{calc} represents, and $\frac{cg}{1+g}\log_2 N$ was derived as an estimate of the communication time per node, similar to T_{comm}. Therefore we can rewrite (7) as

$$f_C = \frac{\frac{cg}{1+g}\log_2 N}{\frac{W}{N}} = \frac{T_{comm}}{T_{calc}} \tag{8}$$

so it follows that the results of Fox's analysis apply to Amdahl's model.

The crucial observation is that the fractional communication time can often be written in the form

$$f_C \approx k \frac{t_{comm}}{F(m) t_{calc}} \tag{9}$$

where k is a constant, m is a measure of the grain size (e.g., the number of points in the subgrid acted on by each node). Note that the machine parameters only appear in the ratio t_{calc}/t_{comm}.

Are there algorithms for which f_C is small? Fox [6, 7] analyzes a number of scientific problems and particular algorithms to compute them and derives expressions for how much computation and communication is done at each node for a given value of m. For many algorithms, the function $F(m)$ is shown to be independent of N, the number of nodes. This implies that high efficiency can be attained on arbitrarily large systems. Some representative results are:

Table 1

F(m)	Problem Class
m	1D Grid point Problems Long range Forces
$\sqrt{m}$	Full or Banded Matrix LU decomposition/eigenvalue determination
$\sqrt{m}$	2D Statistical Physics Sparse matrices from 2D finite element/difference
$m^{1/3}$	As above for 3 dimensions
$\log_2 m$	Fast Fourier Transform

Recall that to get high efficiency f_C must be small; when equation (9) holds, this is equivalent to having a large $F(m)$. This is true of all the entries in Table 1 that have a large enough grain size.

These algorithms and others have been implemented on several different hypercubes by Fox and his group. The actual performance and efficiency observed agreed well with the analysis presented above for systems with up to 128 nodes, the largest available at the time the studies were done.

Summary

The current trends in hardware suggest that it is realistic to establish a goal of several hundred gigaflops performance on a single system within three to five years, particularly if distributed memory architectures with large numbers of processors are used. Shared-memory systems may also scale to such performance, but there is less experience with massively parallel shared-memory architectures. Analyses by Amdahl and Fox, supplemented by empirical evidence, indicate that there are many computations that will achieve high efficiency on distributed-memory systems with very large numbers of processors.

VII. Prospects and Conclusions

There are many commercial products in the advanced computer architecture area. Many of the vendors are aiming at the super-minicomputer and mini-supercomputer market rather than aiming at supercomputer performance levels. The field is currently dominated by small companies that often have only one product line. It appears that most of the major manufacturers do not have products in the high-performance area that are based on new architectures, whether they be multiprocessors, wide-word architecture, or lattice. The exceptions are Floating Point Systems with its T-Series and Control Data with the CYBERPLUS system. IBM has the RP3 project for a system that would range from medium to massive parallelism, but that is a research project, not a commercial product. We are not aware of similar projects within Digital Equipment Corporation, ETA, Cray Corporation, UNISYS, and Gould, to name a few of the major manufacturers. Perhaps this is due to the relatively small size of the supercomputer market. Given that there are already half a dozen companies competing for supercomputer customers with "traditional" machines, there may be a feeling that it would be too risky to introduce radically new supercomputers. A second possible factor for the reluctance to develop advanced architecture supercomputers is the immaturity of the field. It is difficult to determine in advance whether a new architecture has high performance for a wide class of applications and is cost effective. Advanced architectures are filled with potential bottlenecks that dramatically reduce the performance that can be achieved.

The specific paths to high performance in today's commercial advanced architecture computers vary widely. Some, like the CHoPP, feature a few sophisticated and complex processors; others, like the Intel iPSC and NCUBE, utilize many microprocessors with a general-purpose instruction set; a few use very large numbers of extremely simple bit-serial processors. In all cases, high performance is gained through replicating components or processors, rather than through the use of new hardware technology that yields higher speed components. This is true of memory, bus speeds, and basic clock cycle times. On the other hand,

software technology is playing an important role. For several systems, compiler techniques are the key to the feasibility of achieving high performance.

We are beginning to see the expected very wide range of performance that is possible in many of the parallel architectures. The FPS T-Series is an extreme example of this, but many of the others (existing or future) such as the CHoPP, the RP3, the ULTRAMAX, the CEDAR, the NCUBE, and the Intel hypercubes have a wide range of performance. cases the smaller configurations of a particular design are systems of modest performance but the biggest configurations reach fairly high performance. This property of the parallel systems should mean that with parallel architectures it will often be possible to have small to medium-sized systems owned by an individual or small group of people that are exactly compatible with true supercomputer systems, with the only difference being the number of processors in the configuration.

It appears that these innovative high-performance systems that are the topic of this paper are maturing rapidly enough in terms of hardware and software that in a very few years—perhaps only two or three—we will see supercomputer facilities that feature one or more of these systems. These will be prototypical supercomputer facilities. Application software packages for engineering tasks, for example, will not yet be available. Optimizing and parallelizing compilers, program development tools, and debugging aids will probably still be less refined than the corresponding software for today's supercomputers, but they should be robust enough for use in real applications. It is important to note that even now several systems with the innovative architectures provide mini-supercomputer performance at an attractive price and, when dedicated to a small group of users, can tackle "supercomputer-sized" applications.

References

[1] "Gigaflop Speed Algorithm for the Direct Solution of Large Block-Tridiagonal Systems in 3-D Physics Applications", D. Anderson, A. Fry, R. Gruber, A. Roy, UCRL-96034, Lawrence Livermore National Laboratory, January, 1987.

[2] "Microcanonical and Hybrid Simulations of Lattice Quantum Chromodynamics with Dynamical Fermions", D. Sinclair, ANL-HEP-PR-86-121, Argonne National Laboratory, October, 1986.

[3] "Monte Carlo Photon Transport on Advanced Computers", W. Martin, seminar presented at Argonne National Laboratory, January 8, 1987.

[4] "Matrix Algorithms on a Hypercube I: Matrix Multiplication", G.C. Fox, A.J.G. Hey, and S.W. Otto, Parallel Computing 4, p.17, 1987.

[5] "Solving Problems on Concurrent Processors", Geoffrey C. Fox, Mark A. Johnson, Gregory A. Lyzenga, Steve W. Otto, John K. Salmon, David W. Walker, to be published by Prentice Hall.

[6] "Decomposition of Scientific Problems for Concurrent Processors", G. Fox, unpublished paper CALT-68-986, February, 1983.

[7] "Algorithms for Concurrent Processors", G. Fox, S. Otto, Physics Today, May, 1984.

APPENDIX C

SUPERCOMPUTING AND STORAGE

Ken Wallgren
Supercomputing Research Center
4380 Forbes Blvd. Lanham Md, 20706

Abstract

In the late 1990's supercomputers will have computational performance approaching one trillion floating point operations per second. These high performance systems will contain tens to hundreds of giga bytes of internal high speed memory either distributed, centralized or combined to provide the required bandwidth for efficient computation. The impact on the on-line storage and mass storage archive subsystems is estimated from actual user experience with current supercomputers.

The on-line portion of the storage subsystem must satisfy the dual constraint of high capacity and very high transfer rates, while the mass storage portion may have greater capacity with a lower transfer rate requirement. The actual 1990's storage subsystem will be composed of both magnetic and optical disk and tape units. Some of the factors critical to design of the total storage subsystem are examined.

Introduction

Examination of the future role of magnetic and optical disk technology, with respect to supercomputers, requires a frame of reference for plausible projections of capacity and transfer rate. By examination of the future direction of supercomputer performance in operations per second, internal memory capacity and channel input/output transfer rates, an estimate can be made of the requirement for the on-line external storage and mass storage subsystem.

Attempts to define what constitutes a balanced computer systems from a microscopic view[1,2] have resulted in a theoretical understanding of the interplay between processor performance, memory capacity and I/O transfer rates, however, the sensitivity to the application program and the architecture makes it seem impossible to obtain actual system level estimates. Actual system estimates, however, can be obtained by examining the historical values for memory capacity, on-line storage capacity and transfer rates as a function of the processing subsystem performance.

Supercomputer Directions

In the next fifteen years it is expected that supercomputers with peak performances of 100 Giga Floating Point Operations Per Second (GFLOPS) will be in common usage. These high performance computers will be of one or more of the three following types: vector pipeline systems with up to 256 units, large two dimensional array processing system with 256x256 or 512x512 processing elements or very high performance multiple instruction multiple data stream architectures like the hypercube family. Since each of these architectures can reach 100 GFLOPS through the efficient use of the fastest technology that is practical, it is expected that both silicon and Gallium Arsenide integrate circuits will be used.

SPIE-The International Society for Optical Engineering, 30th Annual International Technical Symposium on Optical and Optoelectronics Applied Sciences and Engineering, August 1986, San Diego, Ca.

For the purpose of this paper, it is assumed that all architectures capable of providing 100 GFLOPS of performance are solving a given application job within fields, such as, aerodynamics, weather modeling or computational chemistry. Further, it is assumed that the processor subsystem can be treated as a box which requires volumes of data at high rate to sustain performance related to the application. This implies a sequence of high demand application jobs and not a isolated single pass of an application job for benchmark purposes.

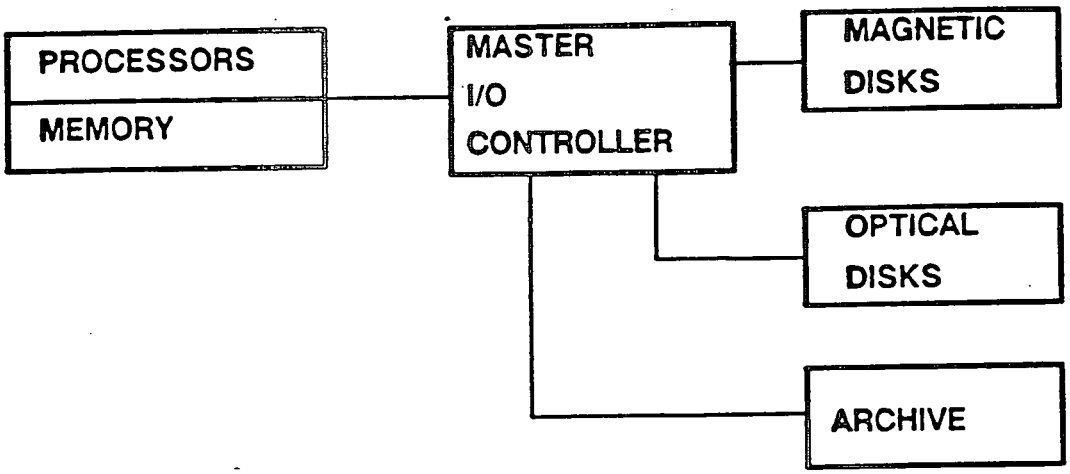


Figure 1. Generalized Architecture

Figure 1 is the generalized assumed system architecture. The processor subsystem and the total memory combine to form what is usually viewed as a supercomputer, all other sections relate to the Input/Output (I/O). The total memory is assumed to hold the program(s), all its data and space for any I/O buffers needed to sustain the computer execution of at least one large application job without additional input during the computation phase. In practice the memory usually stages, at least, one additional job for execution. This is critical to sustaining reasonable system performance. The inclusion of a master I/O Control (IOC) subsystems, addresses the reality that magnetic and optical disks each have a unique features that provide superior performance for a portion of the process.

Balanced Supercomputer Systems

It would be pleasing to have a simple equation or rule of that would let a computer architect known that with a given processor performance a system should have this amount of memory and this amount of storage capacity and this amount of transfer rate to the storage media. The reality is that all of the above parameters are application sensitive. To added to the dilemma, as the computer performance increases the algorithmic methods used within an application change to take advantage of the computing performance available.

An example, in aerodynamics [3] the computer performance of 10 MFLOPS made it possible to change from the Nonlinear Invisicid method, to the Reynolds – Averaged Navier Stokes method for approximating a solution to the Navier Stokes equation which describes the physics of aerodynamics exactly. At 1 GFLOPS the change to the large Eddy Simulation method provides new insights into the flow over aerodynamic surfaces thus allowing the design of more fuel efficient aircraft. At a Tera FLOPS a direct solution to the Navier Stokes equation would be possible for simple structures. For the design of a whole aircraft with this direct solution method it is estimated that a computer

performance of 10**15 FLOPS would be required. Other applications, such as weather and climate modeling, show similar changes in the methods used to approximate the actual physics as computational performance increases.

The above discussion is intended to establish the lack of a exact solution to the problem of prediction of balance in a supercomputer design. The best that can be expected is to get a upper and lower bound on the memory capacity, storage capacity and transfer rate as a function of performance.

Memory Sizing

By examination of existing scientific computers it is possible to bound the relationship of processor performance to memory size and ,thereby, on-line storage capacity. Table I presents a historical examination of the ratio of memory size to performance. The mean of the sample is 0.4, which implies that the new ETA 10 is a reasonably balanced system based on history. This also implies that the CRAY 1A was under sized and that the CRAY 2 is oversized. Experience has validated the that the CRAY 1A was under sized. The new Cray 2, however, may not be oversized. As users and system software designs get experience with the CRAY 2 presumed excess, they may find it critical to efficient usage of this high performance supercomputer. It is only fair to observe that the size of the Cray memory was more likely driven by the availability of 256K RAMS at a reasonable cost, than system balancing concerns.

PROCESSORS	PERFORMANCE SUSTAINED MFLOPS	MEMORY MWORDS (64 bits)	RATIO
CDC 7600	4	0.5	.12
CYBER 203	20	2	.1
ILLIAC IV	25	16	.64
CRAY 1A	30	1	.03
CRAY 1S	30	2	.1
CRAY XMP	80	8	.1
CYBER 205	80	32	.4
Fujitsu VP-200	100	64	.64
Hitachi S-810	126	64	.5
Cray 2	200	256	1.28
NEC	260	64	.25
ETA 10	640	256	.4

Table 1. History Of Memory Size Versus Performance

The memory size of future systems can be estimated assuming that the ETA 10 has a good balance between performance and memory size. The scaling to 10 GFLOPS, 100 GFLOPS and 1000 GFLOPS assumes that as performance increases, the memory size is bounded by M=KP**3/4 and M=P. Where M is memory capacity, K is a constant of about 0.4 and P is processor performance. The M=KP**3/4 bound is valid for most scientific computation such as aerodynamics and weather modeling and the M=KP bound is valid for image processing and most computational chemistry. Schneck [4] has pointed out that the P**3/4 boundary is typical of three dimensions grids where the calculation is performed in time steps, whereas, the M=KP group does not involve time steps. This is consistent with experience at NASA Ames in aerodynamics, weather modeling and computational chemistry.

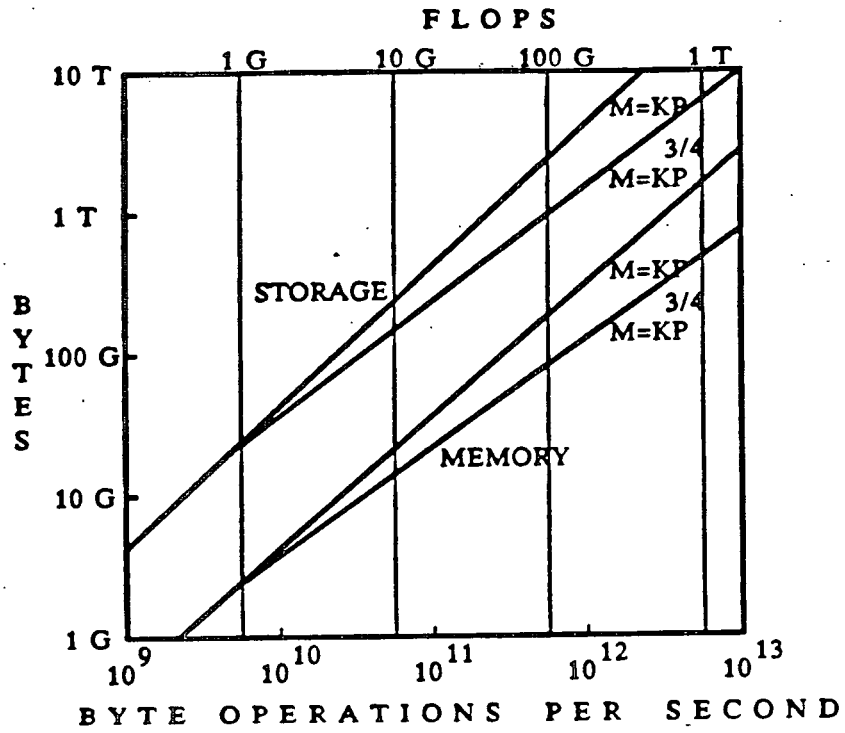


Figure 2. Memory Versus Performance

Figure 2 presents the bounded memory capacity requirement as a function of supercomputer performance. A value of K=0.4 was used from table 1.

On-line Storage Sizing

Applying the experience from CDC, CRAY and LANL [5,6] the ratio of the on-line storage capacity required to memory capacity is 16:1. This provides a reasonable estimate of the disk on-line storage required to support a given computational performance. The data presented in figure 2 at best bound the problem.

The 16:1 ratio is only reasonable when the total system contains an additional mass storage subsystem that is a factor of 16 larger than the on-line storage subsystem. In the case of the ETA 10 and Cray 2, this results in 32 Giga Bytes (GB) of on-line storage and 512 GB of mass storage. It must be noted, that the volume of mass storage is highly sensitive to the application being run at a particular supercomputer site. If very few results are archived than 512 GB maybe to large, however, if the site process large amounts of climate or image data an additional archive of 10's of Tera Bytes maybe required.

Channel Transfer Rates

Existing supercomputer I/O subsystems indicate an adaptation to the availability of supporting peripheral devices. The presence of a number of 50 MB/S channels provides the option of stripping (dividing the data between disks attached to several channels at once)to get aggregate I/O rates of near 100 MB/S has been demonstrated. From a computer architecture and reliability view stripping is not an ideal solution to the transfer rate problem. The

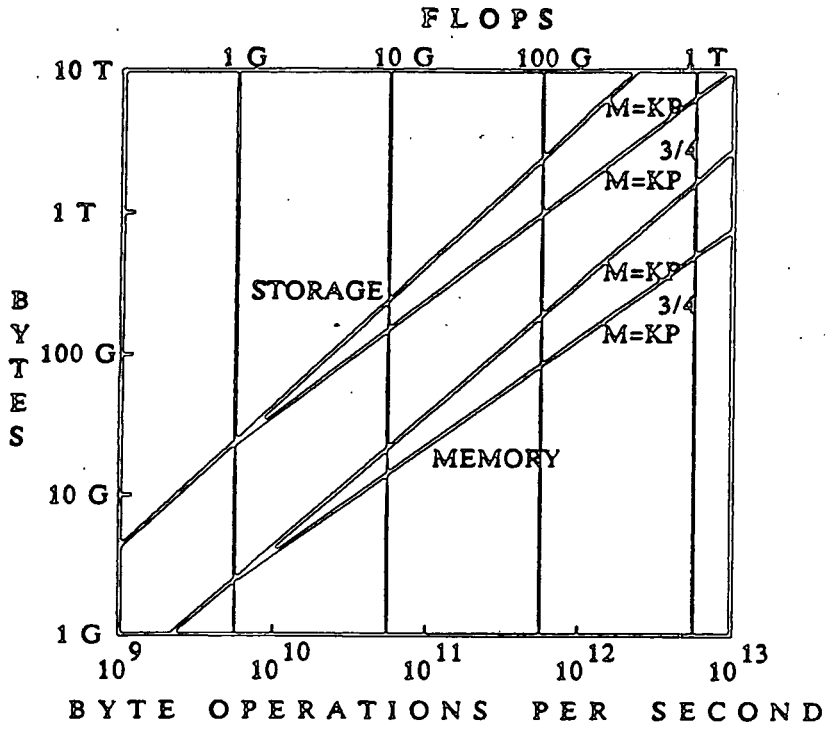


Figure 2. Memory Versus Performance

Figure 2 presents the bounded memory capacity requirement as a function of supercomputer performance. A value of K=0.4 was used from table 1.

On-line Storage Sizing

Applying the experience from CDC, CRAY and LANL [5,6] the ratio of the on-line storage capacity required to memory capacity is 16:1. This provides a reasonable estimate of the disk on-line storage required to support a given computational performance. The data presented in figure 2 at best bound the problem.

The 16:1 ratio is only reasonable when the total system contains an additional mass storage subsystem that is a factor of 16 larger than the on-line storage subsystem. In the case of the ETA 10 and Cray 2, this results in 32 Giga Bytes (GB) of on-line storage and 512 GB of mass storage. It must be noted, that the volume of mass storage is highly sensitive to the application being run at a particular supercomputer site. If very few results are archived than 512 GB maybe to large, however, if the site process large amounts of climate or image data an additional archive of 10's of Tera Bytes maybe required.

Channel Transfer Rates

Existing supercomputer I/O subsystems indicate an adaptation to the availability of supporting peripheral devices. The presence of a number of 50 MB/S channels provides the option of stripping (dividing the data between disks attached to several channels at once)to get aggregate I/O rates of near 100 MB/S has been demonstrated. From a computer architecture and reliability view stripping is not an ideal solution to the transfer rate problem. The

availability of a peripheral device with sufficient bandwidth to support a single applications I/O would simplify system software and allow for efficient staging of jobs.

As in the memory discussion, I/O transfer rates are very sensitive to the application. Classification of applications in to generalized groups provides a basis for estimating the I/O transfer rate as a function of computational performance. Figure 3, from a previous paper [7], is an estimate of the transfer rate required for business, image processing, scientific computing and radar processing as a function of performance.

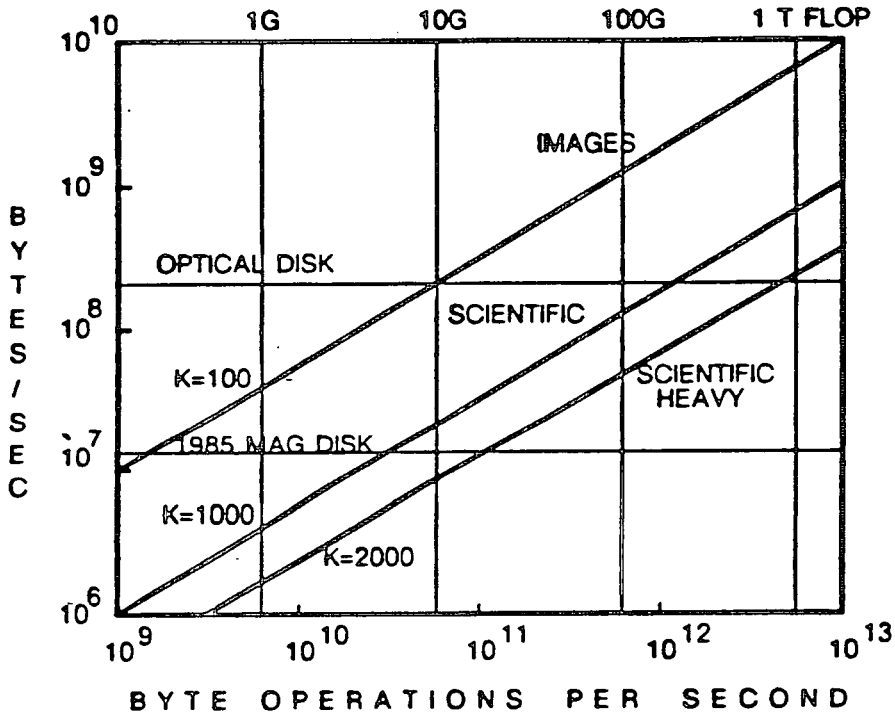


Figure 3. Transfer Rate Versus Performance

Parameters	NOW	1990	1995	2000
FLOPS SUSTAINED	640M	10G	100G	1T
MEMORY SIZE IN BYTES	2G	16G	96G	512G
CHIP SIZE IN bits	256K	1M	4M	16M
# OF CHIPS	64K	128K	192K	256K
STORAGE IN BYTES	32G	256G	1.5T	8T
TRANSFER RATE B/S	50M	400M	1.2G	4G
50% DUMP TIME SEC	20	20	40	60

TABLE 2. SUMMARY OF PROJECTED REQUIREMENTS

The transfer rate and the on-line storage capacity provide the set of boundary conditions for acnieving a reasonable balanced supercomputer system as a function of the performance for a given application group. Table 2 is an attempt

to forecast the memory capacity, storage capacity and the channel transfer rate for systems with performance of 10 GFLOPS, 100 GFLOPS and a Tera FLOPS. The capacity values are the minimum expected for each of the performance levels. The transfer rate values, which historically have been peripheral device driven, were selected on the assumption that large supercomputing applications will have more computations per I/O word.

Storage Options

The challenge for the on-line storage system is to provide the technology that satisfies both capacity and transfer rate boundary conditions of figures 2 & 3 simultaneously. Potentially both magnetic and optical technology can meet the challenge. This section attempts to anticipate the growth in each of the technologies over the next ten years and than to examine how each will satisfy the boundary conditions for anticipated system performance.

A comparison of magnetic and optical disk technology is a comparison of a mature technology and a new evolving technology. At present high transfer rate optical disk technology has not been applied to a supercomputer. Units presently under design and development [8] may provide the first test of an optical disk on a supercomputer. Transfer rates of greater than 100 Mega Mbytes per second (MB/S) will provide new information on the balancing of the I/O to the application.

Optical technology seems to have the capability to provide a single peripheral device that can best provide both the capacity and the transfer rate while retaining a simplicity that will increase the expected mean time to failure. Table 3 is a comparison of what is required in both magnetic and optical technology to provide the on-line storage for a 10 GFLOPS computer system. Similarly, Tables 4 and 5 are estimates for a 100 GFLOPS and a Tera FLOPS computer system. The tables contains estimates of the anticipated performance of magnetic and optical disks in the future based on the historical growth rate of magnetic technology and assuming a comparable growth of optical technology once it is established and tested [9].

Requirements For A 10 GFLOPS System
Capacity 250 Giga Bytes
Transfer Rate 400 Mega Bytes/Second

	MAGNETIC		OPTICAL
	IBM 3380X	IBIS	RCA
CAPACITY EACH GB	10	4	125
TRANSFER RATE EACH MB/S-	6	48	200
NUMBER OF UNITS	67	64	2
TOTAL CAPACITY GB	670	256	250
TOTAL TRANSFER RATE MB/S	402	3072	400
NUMBER OF SURFACES	1072	1024	48
NUMBER OF HEADS	2144	2048	48
ESTIMATED SYSTEM COST $M	11.1	4.1	3.2

Table 3. 1988-1990 On-line Storage

Requirements For A 100 GFLOPS System
Capacity 1.5 Tera Bytes
Transfer Rate 1.2 Giga Bytes/Second

	MAGNETIC		OPTICAL
	IBM 3380X	IBIS	RCA
CAPACITY EACH GB	40	8	250
TRANSFER RATE EACH MB/S	12	48	200
NUMBER OF UNITS	100	188	6
TOTAL CAPACITY TB	4	1.5	1.5
TOTAL TRANSFER RATE GB/S	1.2	9	1.2
NUMBER OF SURFACES	1600	3000	144
NUMBER OF HEADS	3200	6000	144
ESTIMATED SYSTEM COST $M	13.5	12.2	9.6

Table 4. 1991-1994 On-line Storage

Requirements For A 1 Tera Flops System
Capacity 8 Tera Bytes
Transfer Rate 4 Giga Bytes/Second

	MAGNETIC		OPTICAL
	IBM 3380	IBIS	RCA
CAPACITY EACH GB	80	16	500
TRANSFER RATE EACH MB/S	24	100	1600
NUMBER OF UNITS	167	500	16
TOTAL CAPACITY Tera Bytes (TB)	13	8	8
TOTAL TRANSFER RATE GB/S	4	50	25
NUMBER OF SURFACES	2672	8000	384
NUMBER OF HEADS	5344	16000	384
ESTIMATED SYSTEM COST $M	16	25	16

Table 5. 1995-2000 On-line Storage

The advantages of optical technology for high capacity, high transfer rate is related to the simplicity of the device in terms of the number of moving parts and the proximity of the moving parts to the media surface. The typical optical head is 150 times as far away from the active surface as the typical magnetic head. This combined with the number of heads required to achieve the required transfer rate results in a higher expected mean time to failure for the optical subsystem. Reference [10] relates the disk reliability directly to the number of disk accesses (head movements). An increase in the number of heads without a decrease in the total number of accesses should decrease the reliability of the storage unit. This is the case for magnetics.

The reliability consideration must be counterbalanced by the magnetics technologies clear advantage in access time. However, the access advantages is only important if the transfer volume is less than 1 MByte. The higher transfer rate of optical technology can compensate when both access and data transfer times are considered. It is noted, that individual program storage is typically less than 1 MByte.

Future Storage Subsystems

The supercomputer industry has a history of taking advantage of what ever technology is available that will make the total system provide increased cycles to the user. This implies that optical and magnetic technology will be used for there relative advantages. The single most difficult obstacle is for the buyers to understand is that the cost of the peripheral system may be as much as the total main processor and its memory, yet, the cost per MByte of storage is constantly decreasing. The main challenge for the disk industry is to provide cost competitive peripherals that meet the combined high capacity and high transfer rate requirements.

By the mid 1990's the storage subsystem must provide capacities of 8 Tera Bytes and transfer rates of 4 Giga Bytes/second to allow sustained Tera FLOPS performance. Achievement of these goals requires a capacity increase of 8 and a transfer rate increase of 20 over state of the art optical systems planned for 1988 delivery.

References

1. H.T. Kung, "Balanced Computer Architectures", J. Complexity, V.1,1985.
2. H.L. Resnikoff, "Balanced Computer Architectures and Cost Effectiveness" (to be published)
3. V.L. Peterson, "Character of Supercomputer Workload at NASA ARC" NRC Feb. 7, 1986.
4. P. B. Schneck,To be published
5. L.M. Thorndyke, "Supercomputers and Mass Storage: The Challenges and Impacts", IEEE Symposium on Mass Storage Systems, Tucson, AZ. Nov. 1985.
6. R.H. Ewald and W.J. Worlton, "A Review of Supercomputer Installation Mass Storage Requirements" IEEE Symposium on Mass Storage Systems, Tucson, AZ. Nov. 1985.
7. K.R. Wallgren, "Optical Disk and Supercomputers" SPIE Optical Mass Storage Conference, Jan 1985.
8. M.L. Levene, "A High Data Rate, High Capacity Optical Disk Buffer" IEEE Symposium on Mass Storage Systems, Tucson, AZ. Nov. 1985.
9. A.S. Hoagland, "Information Storage Technology a Look at the Future" IEEE Symposium on Mass Storage Systems, Tucson, AZ. Nov. 1985.
10. M.A. Tyler, Hard Facts on Hardware Reliability", Datamation, Oct 15, 1984.

SUMMARY

PARAMETERS	NOW	1990	1995	2000
FLOPS SUSTAINED	640 M	10 G	100 G	1 T
MEMORY IN BYTES	2 G	16 G	96 G	512 G
CHIP SIZE IN bits	256 K	1 M	4 M	16 M
# OF CHIPS	64 K	128 K	192 K	256 K
STORAGE IN BYTES	32 G	256 G	1.5 T	8 T
TRANSFER RATE B/S	50 M	400 M	1.2 G	4 G
50% DUMP TIME SEC	20	20	40	60

ETA Systems, Incorporated
1450 Energy Park Drive
St. Paul, MN 55108
(612) 642-3400

ETA SYSTEMS

APPENDIX D

November 23, 1987

Mr. Norman Kreisman
Advisor, International Technology
U. S. Department of Energy
Office of Energy Research
Washington, D. C. 20585

Dear Norm:

Per our discussion, enclosed is the paper on peripherals subsystems that ETA Systems is submitting. You are authorized to include this paper in your report at your option.

Thank you for this opportunity.

Sincerely,

Lloyd

L. M. Thorndyke
Chairman of the Board

ETA SYSTEMS

THE NEED FOR SUPERCOMPUTER PERIPHERALS

During the past five years, the computer industry has undergone significant changes, not the least of which has been the emergence of an identifiable supercomputer segment. Prior to that time, the entire supercomputer industry consisted of two small American suppliers vying for a worldwide market of 15-25 systems per year.

Radical changes in the architecture, technology and application of supercomputers have led to vast increases in computing power and broader uses. The introduction of these machines to new users and the entry of new competitors, notably three large Japanese companies, into the supercomputer marketplace have increased the CPU performance as these competitors strive for recognition.

The performance of the Mass Storage System (Mass Storage is considered semi-conductor memory, magnetic disk, archival devices such as tape, optical, or new technology and compatible tape) has been sadly neglected and languishing in performance capacity and reliability as applied to supercomputer needs.

The networking of these devices into an efficient Mass Storage System that can satisfy multiple supercomputer systems does not exist now. Further, the technological advances of supercomputers have not been applied to Mass Storage Systems because of economic reasons. There is insufficient volume to recover development costs due to low production rates. Also, any safe Mass Storage System (no risk) technical approach will not yield capacity performance and reliability required. Thus there is even less interest.

In the magnetic disk area, the net effect is to force the supercomputer system to use disks developed for the more traditional low performance systems. These devices are designed with a heavy emphasis on reducing the cost of storage, but are deficient in their ability to transfer data at the rates required by today's supercomputer.

The three most crucial requirements for supercomputer storage are performance (transfer rate), capacity and data reliability. These must be viewed in light of two key characteristics of the current supercomputer systems. First, the individual computers within the system are faster than ever, and second, a large system will contain multiple computers. A reasonable estimate is that the supercomputer demands on disk have grown by a factor of 25 during the past five years but the disks system performance has grown less than three times the next generation of supercomputer devastates this ratio.

It is clear that the pressing demand for supercomputer mass storage disks can only be met by a concerted and dedicated effort aimed at that specific objective. Further, it is clear that such an effort will require the application of human and financial resources for a period of several years. Equally important, the prospect of willing and ready buyers must exist. The benefits of this endeavor will be vastly improved utilization of the supercomputer systems and the ability to apply supercomputer power to problems not currently feasible today.

The archival effort is fragmented and discouraging in progress and again aimed at small computer markets for good reasons cited before. The Government has attempted to directly contract developments critical to their missions. If successful, these efforts must be integrated into the mass storage system that allows efficient systems use.

Finally, the hard copy capability and quality required must be addressed. The significant graphics growth in the next few years will accelerate the need for increased hard copy capacity and quality output. (We still produce reports and memos on work). If we don't plan and emphasize this, we will find another bottleneck has developed.

The industry needs encouragement and shared risk taking to immediately address these problems. We all recognize the problem but can only start the effort if bona fied customer (orders) for prototype devices come forward and risk-sharing money and people are identified.

In the interest of U.S. supercomputer companies' long-term business goals, an avenue must be opened to provide these needed mass storage systems capabilities at the earliest possible time. It is clear that the U.S. supercomputer companies will not be able to depend on the creation of high-technology storage, archive and hard copy systems targeted solely for supercomputing by the general peripheral vendor mileau. Instead, a new generation of supercomputer peripherals will have to be brought into being as true "subsystems", based on technologies employed in the standard "non-supercomputer" marketplace. Only through the creative leveraging of state-of-the-art componentry which is being produced by profitable, successful peripheral manufacturers, is there a practical chance of supporting today's and tomorrow's generations of supercomputers. In order to encourage ventures in this neglected area, we must have prospects of early risk money and orders.

SOFTWARE FOR SUPERCOMPUTERS
— A REPORT —

Prepared by the Subcommittee on Supercomputers
of the
Committee on Communications and Information Policy
United States Activities Board

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS

Revised, 1988

SOFTWARE FOR SUPERCOMPUTERS
A REPORT

EXECUTIVE SUMMARY

A summary of a report prepared
by the IEEE Scientific Supercomputer Subcommittee
of the Committee on Communications and Information Policy.

Early supercomputers had poor software.

The first truly unique supercomputer architectures started to appear in the early 1970s. The Burroughs Illiac IV, the Texas Instruments ASC, and CDC's Star 100 were built in small quantities and the software was not highly developed. In many cases the purchasers provided most of the software themselves. It is important to understand that the very uniqueness that allows these computers to yield very high performance also forces users to expend significantly more effort in optimizing their codes to achieve even a fraction of this potential power. These early machines were very difficult to use and the software was not easily optimized. When the current class of supercomputers (called Class VI machines) started to make their appearance in the late 1970s and early 1980s, new software had to be provided. Because of difficulties in formulating and optimizing the programs appropriately for the newer vector computer architectures, the efficiency of use of Class VI architectures suffered (and still does).

Portability, optimization, algorithms.

We have three main needs associated with supercomputer software: portability, language and compiler related software, especially automatic optimization, and architecture–appropriate algorithms.

Portability of code was and is an important problem.

The ability to take programs from one manufacturer's machines to another or even to move code to later generations of the same equipment is knows as portability. This has been a continuing problem in the computing industry, especially in supercomputing. Lack of portability not only causes premature obsolescence of users' codes, but also shortens the lifetime of system code supplied by the manufacturer, thus making it even more difficult to justify the heavy cost of system code development. Some early government users of supercomputers have tried over the years to provide continuity from one generation to the next. For example, a group at the Lawrence Livermore National Laboratory designed the

Livermore Time Sharing System (LTSS) operating system in 1963 for the Control Data Corporation 1604 computer and continued its development through later CDC computers, including the CDC 3600, 6600 and 7600, thus providing compatibility from generation to generation. The version that runs on Cray computers is now called CTSS and is being used at Livermore and by other DOE laboratories. (It is interesting to note that two of the four newly created NSF Advanced Scientific Computing Centers—at San Diego, California and Urbana, Illinois—have recently opted to use CTSS as well.)

Optimization software is expensive, difficult.

Program optimization is another area of crucial importance to supercomputers. Good automatic optimization is required to achieve a higher percentage of the potential speed of supercomputers, better utilization of scarce manpower, and better portability. A good deal of work is now being done on automatic program optimization. (It is perhaps interesting that very little of this work is being done by the hardware vendors.) But even after using the best of these optimizers, the performance typically obtained from the machine is far less than the peak performance possible.

Better parallel algorithms are needed.

Better algorithms can make a major difference in the feasibility of some applications. One only has to think of Fast Fourier Transforms (FFT) and the Simplex method to recognize the impact better algorithms can have. Algorithms are especially important on supercomputers because they need to be specially designed to take advantage of the vector and multiprocessor parallelism. Fortunately, some minisupercomputers now available make it possible to experiment with new parallel algorithms, and many of these systems are being used in this fashion. However, we are a long way from saying that we know how to use parallel processors efficiently for most problems. We now have the tools to study parallel algorithms and we must make these tools available to the algorithms community.

Newer (multiprocessor) architectures make the problem worse.

To make matters worse, machines with new architectures possessing highly parallel structures are now being designed and built. At the moment, we are exploring the capabilities of high performance systems containing only a few parallel processors. A number of supercomputer systems being planned are somewhat larger, having up to 16 processors. Yet good optimization software does not yet exist even for these low levels of parallelism. Machines with new architectures possessing highly parallel structures including hundreds, even thousands of processors, are now being designed and built. Optimization for these machines promises to be even more difficult and labor intensive than the last generation of machines.

Efforts to design automatic optization software to alleviate this problem is at a very early stage and the costs involved in developing this software are so high and the efforts to develop it are so fragmented, that very little may ever see the light of day.

Lack of coordination.

Many of our difficulties stem from a lack of coordination of effort. Manufacturers were reluctant to cooperate out of fear of antitrust laws, and they were reluctant to finance significant software development for what they viewed as a short term product. When the Japanese entered the supercomputer competition, they took a more global approach to supercomputers and treated them as important, highly marketable entities. Having established supercomputers as a national priority, they were able to take a longer term view of software development for these machines, a policy which is proving to be successful. The performance of their machines shows the results of their foresight by outperforming the U.S. supercomputers in many instances, despite their relatively recent entry into the supercomputer arena. One should wonder how our performance will compare with theirs in a few years.

Recommendations.

There is a need for government directed focus.

It must be remembered that supercomputer software is not just important to the position of the U.S. supercomputer industry in the world market, but it is also crucial to a much broader spectrum of industries that depend on supercomputers for the design of competitive products. This committee believes that the importance of supercomputers in government and industry is just being recognized. Nevertheless, software for these computers remains underdeveloped due to the relatively small size of the supercomputer software marketplace (compared, for example, to the market for workstation and personal computer software) and the fragmented and uncoordinated efforts in this area. Although some attempts have been made to remedy the situation, we believe that it would be in the best interests of the United States if the government were to provide more focus on this problem through the following actions:

(1) Stimulate the supercomputer industry by underwriting some of the costs of hardware and (especially) software development through prepurchase programs.

(2) Improve the state of supercomputer software by direct research and development contracts and grants to industry and government laboratories.

(3) Increase basic research funding in supercomputer software.

(4) Establish a formal coordinating body to better focus existing development efforts through standards for software portability, and to provide interagency coordination of Federally funded research efforts.

(5) Establish several National Supercomputer Software Research and Development Institutes (NSSRDI), as recommended in the SIAM report and in an earlier report from this committee[1] to be associated with existing supercomputer centers, both academic and at national labs. The early successes of some of the present national labs is evidence of the potential for success of such institutes. These institutes should have the following goals.

- Advise the Federal government on matters relating to supercomputing;
- Set common software specifications for supercomputers;
- Carry out practical research in structuring algorithms and applications for supercomputers, including parallel (multiprocessor) algorithms;
- Develop software packages, including operating systems and compilers, that would be suited for a wide variety of supercomputers;
- Devise performance measures for supercomputers; and
- Package these products for government, educational, and industrial use.

---

[1] *A National Computing Initiative: The Agenda for Leadership,* Published by SIAM, Philadelphia, 1987. *Software for High Performance Computers,* Prepared by the Subcommittee on Supercomputers of the Committee on Communications and Information Policy of the Institute of Electrical and Electronics Engineers, December, 1985, Washington D.C.

The IEEE Subcommittee on Supercomputers of the Committee on Communications and Information Policy has produced a number of position papers on supercomputers. For further information, or to be placed on a continuing mailing list, contact:

Heidi F. James
IEEE Washington Office
1111 19th Street, N.W.
Suite 608
Washington, D.C. 20036
(202) 785–0017

TABLE 1A

COMPUTER PERFORMANCE

SOLVING A SYSTEM OF LINEAR EQUATIONS WITH LINPACK
(FULL PRECISION--ALL FORTRAN)

	SYSTEM	MFLOPS
*	NEC SX-2	43
*	CRAY X-MP-4 (1 Proc 8.5 ns)	39
*	NEC SX-1	36
*	NEC SX-1E	32
*	CRAY X-MP-2 (1 proc)	24
	NAS AS/XL V60	21
*	CRAY-2 (1 proc)	18
*	AMDAHL 1200	18
*	CDC CYBER 205 (2-pipe)	17
*	FUJITSU VP-200	17
*	HITACHI S-810/20	17
*	CRAY 1-S	12
	IBM 3090-200/VF (1 proc.)	12
	NAS AS/9160	8.3
	FUJITSU M-380	6.3
	CDC CYBER 875	4.8
	AMDAHL 5860 HSFPS	3.9
	CDC 7600	3.3
	CONVEX C-1/XP	3.0
	FPS-264/20 (M64/50)	3.0
	IBM 3081 K (1 proc)	2.1
	HONEYWELL DPS 8/88	1.7

* SUPERCOMPUTERS
TAKEN FROM TECH MEMO 23 (9/1/87)
Jack J. Dongarra (Argonne National Laboratory)

TABLE 1A (CONTINUED)

COMPUTER PERFORMANCE

SOLVING A SYSTEM OF LINEAR EQUATIONS WITH LINPACK (FULL PRECISION--ALL FORTRAN)

SYSTEM	MFLOPS
AMDAHL 470 V/8	1.5
IBM 370/168 (fast mult.)	1.2
AMDAHL 470 V/6	1.1
ELXSI	1.1
°	
IBM 4381 MG2	.96
°	
DEC VAX 8600	.48
°	
°	
°	
DENELCOR HEP-1	.21
°	
°	
°	
IBM PC (W/8070)	.012

TABLE II

WORLDWIDE DISTRIBUTION OF SUPERCOMPUTERS

(INSTALLED AS OF 12/1/87)

COUNTRY	NUMBER OF SUPERCOMPUTERS
UNITED STATES	146
JAPAN	81
FRANCE	25
GERMANY	24
UNITED KINGDOM	16
CANADA	9
HOLLAND	5
NORWAY	3
AUSTRALIA	3
SWITZERLAND	2
ITALY	2
SWEDEN	1
ABU DHABI	1
SAUDI ARABIA	1
TAIWAN	1
DENMARK	1

WORLDWIDE DISTRIBUTION BY APPLICATION (APPROXIMATE)

APPLICATION:	NUMBER OF SUPERCOMPUTERS
Government	130
Manufacturing	71
Education	68
Process Industries	37
Utilities Industries	4
Others	11

CRAY RESEARCH, INC.

608 2nd Avenue South, Minneapolis MN 55402 • (612) 333-5889 TELEX #290789

February 11, 1987

Mr. James F. Decker
Deputy Director
Office of Energy Research
Department of Energy
Washington, D.C. 20585

Dear Jim:

I am writing in response to your letter of January 7th inviting white papers to be incorporated in the FCCSET Committee's new study of supercomputing in the United States. While I would not put this letter in the category of "white paper", I still hope that my thoughts will be useful in your discussion.

In any case, all of us at Cray are looking forward to seeing your report when it is issued in the early Spring. Your work in the past has had an important impact on the industry, and I would expect no less from your current project.

First of all let me say that in the four years since your committee was first organized, the pace of development in supercomputers has shown no sign of slowing down. This is true both in the laboratory and in the marketplace.

Regarding the latter, all of us associated with the field have been somewhat surprised and certainly impressed by the strong market interest in our technology. Speaking just for Cray Research, last year we signed contracts for 46 supercomputers. This is fully 4 times the number we thought it would be possible to sell in a year at the time we entered the marketplace in 1976. Furthermore, the demand for supercomputers has grown significantly in all three of our marketplaces; commercial, government, and universities. It is interesting to note, in fact, that while Cray Research got its start selling to government laboratories, the number of our commercial customers has exceeded that of our government customers since 1983. Specifically, we closed 1986 with 30 government customers and 50 commercial customers.

I hasten to point out, though, that our government customers are still our most significant clients with typically more and larger systems installed per site.

In the last 3 years, thanks in no small measure to the work of your committee, the university market for our supercomputers has expanded significantly. We ended 1986 with 16 university customers and could add as many as 7 or 8 in 1987. This, of course, has been greatly stimulated by the support of the

National Science Foundation, which in turn was stimulated by the FCCSET Committee Report in 1983. There has been a large multiplier effect from the NSF program as well. In fact, only three of our new university customers have received direct NSF support to set up supercomputer centers. The rest have moved ahead on their own.

In recognition of the importance for the future of the university market, Cray Research is also providing its own specific support. Through this year, we are providing $7.8 million in research grants in support of over 140 projects at 12 universities. Each of these projects is designed to explore how supercomputers can best be used to push forward the boundaries of knowledge in many scientific fields.

In our own laboratories, as I indicated earlier, supercomputer technology itself is still moving rapidly. Since your last report, we have introduced the Cray-2 and installed it at a half a dozen customer sites. Also, we have improved the performance and lowered the cost of our mainline X-MP series.

In the new product field, we expect to introduce a successor product to the X-MP this year which should show at least a threefold improvement in performance.

Seymour Cray's Cray-3 project has now entered the preproduction phase; and we hope to demonstrate a meaningful prototype sometime in 1988. Also, Steve Chen has initiated his "MP" project and has established the basic system design as well as the underlying technology he will use to construct a machine.

In software, we have introduced a new operating system across our product line called UNICOS and a new FORTRAN COMPILER called CFT-77. Last year the number of major applications specifically adapted to the Cray architecture topped 400 for the first time.

With all this progress, problem areas and opportunities for improvement still exist. For example, while access to supercomputers in the academic community has certainly been improved, it is still true that many researchers cannot get all the supercomputer time they need. This is evident from the number of requests we have had to support supercomputer research in our own grant program.

Furthermore, the support that has been provided cannot be considered a one time shot in the arm. New technology is coming constantly and needs to be introduced into new and existing supercomputing centers. I am specifically concerned that NSF has shied away from supporting Cray-2 technology at the University of Minnesota because of a concern that it was "unproven" when it was first installed there.

Our greatest challenge, however, comes from overseas and especially across the Pacific. Competition is becoming stiffer. Initial readings on the new NEC machine are impressive, and I cannot believe that Fujitsu and Hatachi are

not at work on new generations of their own equipment. Obviously, we have no alternative but to move even faster.

This task of staying ahead is not made easier by a concentration of semiconductor technology in Japan. We at Cray are very concerned about having to buy our parts from our strongest competitors. This concern is best expressed in a memo from Don Whiting which I have attached. Don, as you know, directs all of our manufacturing activities. He has also made several trips to Washington to voice our concerns.

Finally, newly industrialized countries are becoming an important factor in our field. Over the next few years, we see the possibility of selling 25 or more supercomputers to various countries such as Korea, Taiwan, Singapore, China, India, Brazil, and others. Obviously, this important business for us; but it is even more important for our competitors. One of the greatest assets we at Cray have is our hundred or so customers that are constantly telling us how to make our machines and software better. Our competitors frankly do not have this rich resource; and the newly industrialized countries could present them an opportunity to create such a resource.

The biggest problem we have in developing that marketplace ourselves, is export control. Clearly, there are important security interests to be considered and preserved in dealing with the newly industrialized countries. At this time, however, we are dangerously close to developing an adversary relationship between U.S. suppliers and our own government in trying to deal with this challenge and opportunity. Any help your committee could be in casting light on this problem would be greatly appreciated and highly productive.

Jim, these are some of my thoughts presented in a fairly informal fashion. In spite of their informality, I hope they are helpful. If you would like more information or further elaboration, please do not hesitate to contact me. I would happy to meet with your committee, as I have in the past, at anytime.

I hope we meet again soon. In the meantime, best regards.

Sincerely yours,

John A. Rollwagen
Chairman and CEO

JAR/pmr

CRAY RESEARCH, INC.

INTEROFFICE MEMORANDUM

TO: John Rollwagen DATE: February 2, 1987

FROM: Don Whiting

SUBJECT: FOREIGN DEPENDANCE - SUPERCOMPUTING

In the late 1970's and very early 1980's, Cray Research, in addition to many other companies, discovered that high-tech I.C.'s were Japanese as opposed to American. The 64K DRAM was the principle part that created the revolution in general. For Cray, it was the 4K x 1 ECL RAM. Every leading edge memory part has been bought from Japan since that switch we made in 1980.

In addition to having advanced technology, the quality and reliability of the Japanese parts have been superior by at least an order of magnitude for six to seven years. In 1980 the domestic I.C. companies launched significant programs to improve quality and reliability which have been reasonably successful. While our Japanese suppliers still lead the quality race, the domestic people are quite close. Progress has been made.

Our Japanese suppliers, while performing very well, are also our competitors. They have a stated objective to dominate the computer marketplace and specifically targeted is the supercomputer area of that market.

With this slight background, some of my concerns with foreign (Japan) dependance and foreign take over of domestic I.C. capability are as follows:

- Whoever controls the I.C. technology will clearly control all high tech end items.
- U.S. I.C. suppliers do not appear to have equal opportunities to sell in Japan.
- If the best technology company is also one of our best competitors, Cray must be careful about how we share our future design plans.
- If we work with competitors, we have to assume we will not get the latest technology first.
- With our unsettled trade policies, we have to assume that I.C.'s from Japan may be restricted, quoted, taxed, dutied, etc... at some point in time.

CRAY RESEARCH, INC.

INTEROFFICE MEMORANDUM

- The best computer companies in the world are also the best most advanced I.C. companies in the world. NEC, Hitachi and Fujitsu are all in this category. IBM, DEC and to some extent AT&T are in this category. We can not buy I.C. technology or even parts from IBM or DEC. AT&T appears to be in a state of change or retrenchment that may also restrict our ability to get parts.

- To date, our best allies in real development have been Motorola and Fairchild. We are beginning to see smaller companies like Gigabit and Performance Semiconductor filling a niche market.

- Joint efforts like MCC do not appear to work well in the United States.

It appears there needs to be a shakeout and consolidation in the I.C. world. A number of large survivors will emerge that will control the merchant market. How this occurs will determine the international flavor of the electronics market. If competition were fair and equal in all countries, the U.S. would still be a dominant force. Given that this does not currently exist, some U.S. Governmental intervention appears necessary. Giving away our semiconductor business will surely result in a loss of control of all electronic products.

Dan

DFW:lm

cc: Les Davis

CONTROL DATA

1201 Pennsylvania Avenue N.W.
Suite 307
Washington, D.C. 20004
202/789-6517

Lolo D. Rice
Senior Vice President
Government Affairs

October 5, 1987

Mr. James F. Decker
Deputy Director
Office of Energy Research
Department of Energy
Washington, D.C. 20585

Dear Mr. Decker:

At the suggestion of Lloyd Thorndyke, I wish to submit some recent recommendations of Control Data/ETA Systems to U.S. policy makers on enhancing the competitiveness of the U.S. supercomputer industry. You will note that some of these proposals were contained in the White Paper we presented to the Federal Coordinating Committee on Science, Engineering and Technology (FCCSET) in February 1987.

We believe that if the U.S. is to retain its world leadership in the design, development, manufacture and application of supercomputers we must recognize the gravity of the Japanese competitive threat and implement appropriate policy responses.

The recent U.S. government investigation of Japanese trading practices in supercomputers apparently concluded that the government of Japan has mounted an aggressive strategy to dominate the global market for supercomputers. This strategy includes protecting the home market against access by American vendors; predatory pricing of supercomputers in the U.S. and third countries; and subsidizing technology development in advanced computing. Such activities are increasing Japan's world market share in this critical sector and could repeat the semiconductor scenario with even more ominous consequences for the national security of the U.S.

Control Data/ETA Systems has already set forth detailed recommendations for addressing the question of trade practices in supercomputers in its February 26, 1987 submission to the U.S. Trade Representative. (See enclosure.) In summary, we recommended an ongoing rigorous dialogue with Japan to achieve a level playing field in supercomputer trade. We urged that this dialogue include the whole range of issues from market access to predatory pricing practices. We did not advocate protective tariffs or government subsidies to prop up the domestic supercomputer industry.

We were pleased with the August 1987 exchange of letters between Ambassador Yeutter and Ambassador Matsunaga which appears to open up Japanese government agency procurements through a more transparent process of bidding rules. And we understand that USTR is committed to treating predatory pricing, particularly the so-called "blockbuster" marketing give-aways to universities, as an outstanding unresolved issue in the overall bilateral relationship. Our government must exert pressure to halt predatory marketing practices in U.S. universities.

But a defensive policy alone will not promote the global competitiveness of the American supercomputer industry. Our country requires a proactive, affirmative U.S. government policy of supporting technological excellence in supercomputing. This policy must take at least four forms:

One, the U.S. government should promise system procurements upon successful demonstrations of design goal achievements. A program of guaranteed procurements of supercomputers that satisfy design and performance standards, not unlike that practiced in our fighter aircraft program, would be a powerful incentive to technological preeminence.

In this connection, we strongly urge the Department of Defense, NASA and the intelligence agencies to recognize the profound threat of a dependency on imported supercomputer technology and adopt a policy of domestic content in such procurements. Indeed the GATT procurement code contemplates a domestic reservation in national security or defense industrial base procurements. This policy would of course apply to components and peripherals used in these technologically advanced systems.

Two, we recommend that the government establish a formal program of assigning promising supercomputer design proposals to specific government laboratories and agencies which will procure and integrate these systems into their working environments. Such cooperation would naturally focus on software development and applications.

Three, since increasingly supercomputers are product families with a significant range of performance from smallest to largest, we need to relax and simplify export control procedures. Even where a standard set of safeguards is required, the interagency processing of export licenses simply takes too long and does not seem to consider adequately differing performance ranges. If our government cannot match Japanese license processing time, the American supercomputing industry will not achieve competitive production levels. Improving the licensing process will also demand far improved communications between industry and government officials.

Four, NSF should give broader support to U.S. university procurements of U.S. supercomputers. It is a national disgrace that today there are more supercomputers in Japanese universities than in our own universities.

We thank you for still another opportunity to submit our views on government policy for supercomputing and look forward to the final report and recommendations of FCCSET.

Sincerely,

Lois D. Rice

Lois D. Rice

LDR/jg

CC: R. M. Price
T. C. Roberts
L. M. Thorndyke

Control Data Corporation/ETA Systems

COMPETITIVENESS IN SUPERCOMPUTER TECHNOLOGY

February 26, 1987

Control Data Corporation/ETA Systems

COMPETITIVENESS IN SUPERCOMPUTER TECHNOLOGY

I. SUMMARY

The Section 305 investigation will help focus attention on the critical importance of the U.S. supercomputer industry to our national security and economic well-being. This inquiry should also enhance understanding of the direct link between the supercomputer industry and U.S. technological competitiveness and the central role of supercomputers in driving technology all across the domestic industrial front. Moreover, the 305 investigation is an important first step toward future negotiations with the government of Japan on a whole range of issues addressing concern over Japanese practices in supercomputer trade.

CDC/ETA neither favors the imposition of tariffs or quotas to protect the U.S. supercomputer industry, nor advocates any specific trade law response to Japanese trading practices in supercomputers. The U.S. government is in a far better position to determine an appropriate course of action under U.S. trade law.

CDC/ETA nevertheless believes that present trends in supercomputer trade, together with evidence suggesting certain unfair trade practices, require the U.S. government to initiate a more coordinated and rigorous dialogue with Japan to assure a level playing field in supercomputer trade and fair market access. This dialogue should include issues such as R&D, achieving equity in technology flows and pricing strategies in the U.S. and third country markets.

Through government-to-government negotiations, U.S. trade officials will hopefully be able to persuade the Japanese government to take the necessary steps to ensure that its practices represent an even-handed, fair and equitable approach to competition in the supercomputer industry. In conducting these negotiations, U.S. trade representatives should view Japan as a strategic ally and economic partner of the United States. Technological cooperation must replace adversarial trade practices in future trade relations.

At the same time, the U.S. government must recognize that Japanese trade practices are deeply entrenched and that America is losing leadership to Japan in a number of critical, leading-edge industries and technologies, including supercomputers. Supercomputers help drive the technology throughout the computer industry, have direct

effects on the infrastructure that supports supercomputers, such as semiconductors, and provide computational ability to solve problems in other fields. The loss of technology leadership in this sector would increase U.S. dependency on foreign sources of supply.

CDC/ETA does not advocate government subsidies to prop up the domestic supercomputer industry. But, in the interests of the national security of the U.S. and the quality and productivity of America's industrial infrastructure, the U.S. government must act affirmatively to encourage innovation in new high technology industries and promote a viable broad-based American supercomputer industry.

CDC/ETA urges the U.S. government to adopt a policy of proactive support for the still emerging domestic supercomputer industry through "commerciality," i.e., the competitive process of federal government commercial product procurement. This policy would promote the commercial development of critical American industry and technology as well as R&D excellence. In addition, the U.S. government must implement tax incentives to encourage research and development, promote research in educational institutions and excellence in education, enhance the capabilities of the National Science Foundation to support science, technology and advanced computing and expedite the processing of export license applications while limiting the scope and burden of export controls.

This paper suggests several approaches for U.S. government consideration in support of the domestic supercomputer industry. The point is that trade law remedies alone will not suffice to maintain U.S. competitiveness in supercomputer technology.

II. BACKGROUND

During the 1980's supercomputer* applications expanded to more and more commercial and industrial uses. Of the approximately 245 supercomputers worldwide, 180 have been installed since 1982. Present estimates are that by 1990 this market will expand from the current $500M annually to $1.5-2.5B.

To generate profits, supercomputer companies must design quickly and at a time when the design cycle is rapidly compressing. To survive, companies looking at the 1990 market must get a reasonable market share to sustain both the design stages of today and the faster designs and lower costs of tomorrow. In short, supercomputer companies will exit the marketplace unless they can support required research and development costs.

Supercomputers were once an exclusively American industry. Control Data Corporation (CDC) developed the world's first supercomputer more than 20 years ago. CDC and Cray Research, another Minnesota based company, dominated the marketplace until the mid 1980's when in just one year, 1985-86, Japanese companies increased their orders by 100% -- from 30 to 64. The government of Japan and the three industrial giants, Fujitsu, Hitachi and NEC, now represent a serious threat to U. S. leadership in supercomputers.

The real issue, however, in supercomputer competitiveness is technological leadership. The United States has been a world leader in harnessing the creativity of its engineers and scientists, developing new and innovative designs for supercomputers, and commercially exploiting the business opportunities created by those innovations. For the United States to maintain its role as a world innovator, its industries must remain productive at the leading edge of technology.

Leadership in supercomputing - a linchpin industry crucial to our entire industrial base - is important to the military and economic security of the United States and its allies. Leadership is essential to the advancement of scientific knowledge and know-how, and to the ability of American

---

* At any given point in time a supercomputer represents the leading edge of computer capability. Currently, a supercomputer is a computer system capable of solving large scientific problems characterized by a trillion calculations using 1 billion words of data within a 24 hour period. Advanced, highly integrated semiconductor chips, high performance magnetic recording devices, and new forms of software that exploit "parallel processing" techniques, are all essential components in today's state-of-the-art supercomputers.

engineers, scientists and researchers to tackle and solve the most difficult computational problems, involving such diverse applications as weather forecasting, petroleum exploration, weapons research, biotechnology and pharmaceutical research, automotive and aircraft design.

The U.S. has lost its consumer electronics industry; the semiconductor industry may also be lost. If the supercomputer industry follows, stagnation will occur and the U.S. will lose both the leadership in advanced computing technology and the ability to compete in many industry sectors. The price of U. S. failure in supercomputers is therefore extremely high: it is not only the loss of other leading edge technologies but also America's industrial competitiveness.

III. TRADE PRACTICES IN SUPERCOMPUTERS

Targeting

Export targeting involves a combination of coordinated actions intended to enhance the international competitiveness of a specific firm, industry or group of companies. Foreign industrial targeting practices can have an injurious impact on the viability and competitiveness of U.S. industries. This has been evident where the foreign government has sought to develop a particular industry by creating a relatively risk-free environment to provide a competitive advantage the industry would not otherwise have under normal market conditions.

Targeting is different from other potentially trade-distorting practices in that it involves a combination of actions, any one of which may have a marginal impact on the industry's competitiveness, but which taken together artificially create a comparative advantage for the selected industry. Targeting practices typically include certain non-tariff barriers to curtail foreign access to the home market; technology design and development subsidies and research cartel arrangements; and predatory practices designed to penetrate the U.S. market and preempt U.S. exporters in third-country markets. In addition foreign governments often direct private capital as well as government financial resources to the particular industry on a preferential basis, establish an exclusive industry cartel and provide preferential sourcing of government procurement to domestic vendors. All of these practices are intended to provide special protection during the establishment and development of the industry and to promote the successful penetration of foreign markets.

The Japanese government has been under investigation for the type of practices mentioned above in a number of industrial sectors, including automobiles and semiconductors. Now Japan is under scrutiny regarding the supercomputer

industry, with strong evidence of implementing a combination of policies intended to dominate worldwide supercomputer manufacturing.

Responding to such adversarial policies is difficult, however, because damage in the form of lost market share at home and abroad often occurs before appropriate remedies can be implemented. The combination of protected home markets, subsidized technology and predatory pricing in the U.S. and third markets can quickly erode the competitiveness of U.S. suppliers. Reduced volumes in turn force U.S. manufacturers to incur higher unit costs, while insufficient sales revenue and profitability foreclose essential investment in R&D. Unable to commercially sustain the R&D necessary to remain competitive -- particularly in high technology industries where product design cycles are so compressed -- companies inevitably must concede the marketplace to the targeting enterprises.

Thus, targeting policies aimed at critical leading-edge technologies should be addressed when they constitute a threat of injury. Because the supercomputer industry supports the leading technology in the computer industry, competitiveness generally in information processing presupposes a strong domestic supercomputer industrial base. Moreover, if the U.S. becomes dependent upon Japan as the predominant supplier of leading-edge computer technology, the U.S. will lose the driving force in semiconductors and related technology and computational power in other fields.

Market Access

The Government of Japan is currently engaged in a number of practices in the supercomputer industry which are cause for concern. For example, access of U.S. supercomputer vendors to the Japanese home market appears severely restricted as a matter of government policy. While industry analysts have long rated U.S. supercomputers as far superior to the Japanese competition, only six such units have been sold in Japan, and none to a government agency.

It has been widely reported that seven years after establishing a sales and support office in Japan, one leading U.S. supercomputer company had sold only six out of twenty-two supercomputers operating in Japan. Domestic companies like NEC, Fujitsu and Hitachi seem clearly to have had the inside track when government agencies have solicited bids for supercomputer contracts.

When Control Data expressed an interest in a supercomputer procurement announced by the Japanese meteorological agency for weather observations and forecasting -- an area where Control Data computers have been especially competitive worldwide -- the agency insisted that the proposal be submitted in the Japanese language and in only thirty days.

Since the announcement and technical specifications for the procurement were printed only in Japanese, they could not be translated and a response rendered in this short timeframe. Control Data was unable to submit a proposal due to the terms of the RFP.

The same pattern has characterized supercomputer procurements by the Japanese Education Ministry. There appear to be no formal announcements or notifications of upcoming procurements so American supercomputer vendors are effectively precluded from participation. This approach stands in stark contrast to NSF supercomputer procurements, for example, which permit full and open competition and supportive linkages with U.S. universities -- such as in the 1985 Fujitsu/Amdahl-University of Michigan proposal for a federally-funded supercomputer center.

Government Support for R&D

Meanwhile, Fujitsu, Hitachi and NEC, supported by Japanese government R&D subsidies and cartelization policies, continue to develop advanced supercomputer technology. Indeed, the U.S. International Trade Commission concluded as far back as 1983 that the Japanese government had sponsored a number of joint R&D projects involving the fourth and fifth generation advanced computers -- projects from which U.S. companies were excluded. It appears that the Japanese government has not only tolerated but actively encouraged and financially assisted the giant vertically-integrated electronics companies that manufacture supercomputers to establish a number of horizontal ties with respect to research and development in advanced computing. Unlike the U.S. Microelectronics and Computer Technology Consortium (MCC) and other similar American R&D joint ventures in advanced computing which have received no governmental aid for research, the Japanese R&D consortia in supercomputing appear to receive substantial governmental subsidies.

Inequitable Access to Technology

Moreover, the results of U.S. Government supported research and development have been fully available to Japanese and other foreign participation, while government-assisted R&D in Japan has been almost completely closed to American companies and researchers. In the U.S. Japanese companies have had access to the over $18 billion per year of federally-assisted non-defense R&D conducted in the federal laboratories and universities and licensed to foreign applicants. American companies, however, with the sole exception of IBM, have been denied access to computer patents held by MITI on the basis of Japanese government subsidies for R&D in the computer industry.

Finally and equally important, U.S. companies have not been permitted to sponsor and therefore have access to research in Japanese universities or to take equity positions in

innovative small Japanese firms. Japan, on the other hand, has exploited fully both of these opportunities in the U.S. Japanese companies sponsor advanced computing research at a number of leading research universities in the U.S. and have acquired substantial U.S. computer technology by investing in U.S. companies. Fujitsu's investment in Amdahl and the proposed take-over of Fairchild Semiconductor are typical of Japanese technology acquisition in advanced computing.

It may be argued that the U.S. Defense Department research support is in effect a subsidy for U.S. supercomputing. However, it is generally accepted by analysts that the commercial sector in supercomputing is at the leading edge while defense applications are now utilizing trailing edge technology. DoD support of supercomputing for commercial, non-weapons technology development is almost non-existent.

Control Data concludes that in the field of electronics technology, an alarming inequity characterizes the international flow of technology between the U.S. and Japan. Because of this mismatch between Japan's almost unlimited access to American electronics technology and Japan's near total denial of U.S. access to similar Japanese advanced computing technology, together with substantial, direct and exclusionary Japanese government support for its supercomputer industry, the U.S. may be on the verge of losing its comparative advantage in advanced computing technology.

The availability of semiconductor technology related to supercomputers is crucial to the viability of the U.S. supercomputer industry. Supercomputer development requires the newest, fastest, and most integrated technology, but Japanese semiconductor suppliers cannot be counted on to be reliable and timely suppliers of supercomputer technology to the U.S. In fact, if a United States dependency on Japanese components for supercomputers continues to evolve, it is quite probable that advanced technology computers would eventually be supplied only from Japan.

An example of Japanese selective supply occurred in the Autumn of 1984 when ETA SYSTEMS specifically asked Fujitsu to supply a 64k SRAM. Fujitsu indicated that this device would not be available until late 1985; more probably 1986. Within five months, Fujitsu stated that not only would they supply such a device, but also several thousand devices were immediately available. Concurrently, in Japanese trade journals, Fujitsu announced the successful development of a 256k SRAM.

Recently, Control Data engineers identified a particular semiconductor gate-array technology of critical importance to high-performance supercomputers. Japanese vendors informed Control Data that the technology would not be made available to Control Data. This type of incident causes

great concern as to the real motivation behind the Japanese reluctance to supply this technology to a U.S. company.

Pricing Practices

Turning now to the area of pricing strategies, there have been widespread allegations of huge price discounts by Japanese companies to penetrate foreign markets, including the U.S. The U.S. Department of Commerce, for example, has recently scrutinized a transaction in which Nippon Electric Company (NEC) supplied its supercomputer to the Houston Area Research Center (HARC). Estimates are that the transaction price of the NEC SX-2 selected by HARC was in the range of $9 million, substantially below its list price of $22 million. Under such circumstances legitimate questions of below cost sales are raised.

In late 1986, Control Data had a second experience of aggressive pricing strategies by Japanese supercomputer companies. Control Data bid a Cyber 205 supercomputer in a U.S. Air Force procurement for logistics command at Scott Air Force Base. Honeywell Information Systems bid a high-end mainframe computer manufactured by Nippon Electric Company. The Air Force awarded the procurement to NEC because the price was $8 million below CDC. While it is difficult to determine all the circumstances which accounted for this pricing, this type of practice tends to support inquiry into Japanese companies' trading practices.

With specific reference to supercomputers, Japanese policies may also infer a new and more concerning element: downstream dumping. The large, integrated Japanese supercomputer companies each were named parties in the recently concluded Semiconductor Consent Agreement. In that accord, seven Japanese companies agreed to a complicated settlement to end dumping memory chips in the U.S. market to avoid certain imposition of U.S. antidumping duties. Semiconductor chips involved in that judgment would appear to be incorporated by the same Japanese companies in supercomputers marketed in the U.S. and third country markets. Downstream dumping, therefore, may now have emerged as a new concern regarding Japanese trade practices intended to eclipse American supercomputer companies.

Not only in the U.S. and Japan but also in third countries U.S. companies are being confronted with aggressive Japanese trading practices. Worldwide orders by Japanese companies increased from 30 in 1985 to 64 supercomputers by mid-1986 -- an increase in one year of over 100%!* By the end of 1986 the Japanese share of the global supercomputer market of 245 units exceeded 25 percent with concentrated focus on

---

* Japan Economic Survey, January 1987, p.4.

the U.S. market. It is fair to conclude that the practices cited above had to play a significant role in the recent success of these companies. Therefore, the U.S. Government should provide an affirmative response now in order to ensure "commerciality," a "level playing field" and access to the Japanese market and technology, all of which are critical for the survival and success of the U.S. supercomputer industry.

IV. THE NEED FOR A U.S. GOVERNMENT RESPONSE

If the U.S. Government concludes at the end of its Section 305 investigation of Japanese trading practices that there is sufficient concern regarding targeting of the U.S. supercomputer industry, the U.S. trade negotiators should initiate a vigorous dialogue with the Government of Japan to restore and maintain a level playing field in supercomputer trade worldwide.

The focus of U.S. Government policy should be to neutralize any adverse effects of Japanese trading practices. Priority U.S. trade negotiating objectives would include: improving market access in Japan, particularly for government agency procurements; securing U.S. company participation in Japanese Government-funded R&D cooperatives and access to advanced computing technology on a reciprocal basis; and in general, obtaining Japanese Government support for ending unfair trade practices in supercomputers, in the U.S. market and in third countries.

In addition, the U.S. government should move expeditiously to encourage the development of advanced computing technology in the U.S. and proactively support the commercial development of the domestic supercomputer industry. The U.S. must encourage the development and R&D excellence of its critical, leading-edge industries. The government needs to recognize that the U.S. may have already lost leadership to Japan in a number of industries and technologies -- particularly those like semiconductors that support the technology in supercomputers.

Accordingly, the procuring agencies of the U.S. government, including DoD, NASA, DoE, NSF and related programs, must support the domestic supercomputer industry in their competitive procurement policies.

The latter recommendation could be implemented by new legislation or regulations recognizing the importance of supercomputers for U.S. national security and the industrial infrastructure and the potential for a distortion in supercomputer pricing due to strong Japanese government involvement. Specifically, U.S. government procurement law should be amended to require federal agencies procuring supercomputers to take special account of (i) the effects on the domestic industrial base of reliance on foreign-sources

of supply for supercomputers and (ii) the possibility of foreign- source dependency for supercomputers.

In addition, the U.S. government, through the agencies using supercomputer systems, should establish a formal program whereby promising new supercomputer designs are assigned to specific government laboratory/agencies with the responsibility to integrate the systems into their working environments. We believe that the state of the art is rapidly advanced as a result of a joint endeavor. This approach is recommended in several different government laboratories/agencies, including those of the Department of Defense and the National Security Agency. Recognizing that the early introduction of new systems is generally related to software issues that the government has particular expertise in, and to hardware issues that the supplier has particular expertise in, a cooperative arrangement can accelerate the introduction of these new systems to the optimum.

Another recommendation is to have the U.S. government provide system procurement credits for U.S. manufactured components and peripherals.

The components and peripherals used in these technologically advanced systems are key ingredients in the early development of supercomputers, and ultimately find their way into the rest of the computer and electronics industry. The government, therefore, should find ways to encourage the use of domestic components and peripherals in the design and manufacture of supercomputers. We would suggest that a procurement system that provides substantial credit for the use of U.S. made components of the system would encourage their use.

The peripherals industry itself is threatened by Japanese trade practices. While the U.S. government has responded to the need to encourage high performance semiconductor development through such programs as the VHSIC program, it has completely overlooked the fact that for some applications the peripherals, namely the disk drives, limit the performance of supercomputers. The mainstream market for disk drives is not going to provide, for example, the very high data transfer rates required by these applications. Therefore, as a corollary to this recommendation that U.S. peripherals be used, the government must be prepared to increase its support of the magnetic recording industry.

We believe the U.S. government must guarantee procurement of new supercomputers meeting general design goals. The single most important inducement that can be provided to bring new competitors into the supercomputer arena is the promise of system procurements upon successful demonstrations of design goal achievement. Because of the uncertainties in accomplishing specific technology levels, and the desire on

the part of the user community to ever increase the performance of these systems, a cooperative approach is desirable. History has taught that a joint effort by user and vendor generally overcomes initial short falls, especially software, in the most expedient manner.

Finally, the U.S. government must implement tax incentives to encourage research and development, promote research in educational institutions and excellence in education, enhance the capabilities of the National Science Foundation to support science, technology and advanced computing and expedite the processing of export license applications while limiting the scope and burden of export controls.

APPENDIX

1. SUPERCOMPUTER INDUSTRY PROFILE

Supercomputer Applications

Supercomputers have a number of applications critical to U.S. national security and the industrial infrastructure. National security applications were among the first uses for supercomputers and are still a large market segment. Moreover, supercomputers are indispensable for the solution of problems involving weapons design, nuclear effects, naval nuclear reactor design, cryptology/cryptography and space exploration. The deployment of the SDI system will require supercomputers as well.

Other applications important to both national security and civil/industrial goals include meteorology and weather forecasting, both for military and civilian needs. Aircraft design, wind tunnel simulations and structural designs are all highly dependent on supercomputer applications. Much of the research on computational fluid dynamics can be used for civilian and military purposes.

Finally, supercomputers have important commercial, industrial and educational research applications. For example, supercomputers were first used in the petroleum industry for seismic data processing and oil/gas reservoir simulation. Since the beginning of this decade other industrial applications began to expand rapidly so today they now exceed by a considerable margin the supercomputer demands of government and researchers.

The automobile industry has turned to supercomputers to help maintain worldwide competitiveness, with each of the Big Three installing a supercomputer for auto design and crash-worthiness simulation. Another area of considerable growth is electronic computer aided design (ECAD), fueled by the rapid rise in electronic circuit complexity and the sophistication of current computer designs. It is an axiom that it takes a supercomputer to design a supercomputer. The computing power and large memory of the supercomputer are also being applied to graphics, movie animation. After a hiatus of several years because of the lack of Federal support, supercomputer applications in education and research are growing, particularly in the fields of molecular modelling for chemistry and biology.

Manufacturers

There are presently two U.S. supercomputer manufacturers -- CDC/ETA and Cray Research. In Japan only NEC, Hitachi and possibly Fujitsu manufacture supercomputers. These are the only supercomputer companies in the world.

Parts and Design Suppliers

In contrast to the vertically integrated Japanese vendors, the U.S. supercomputer companies are relatively small and must depend on an infrastructure of suppliers for components, peripherals and ancillary equipment.

Cray Research - It has been reported that Cray Research buys most of the logic and logic components for current production models from Japanese sources. It is not known what the source of the design software is, but there is likely Japanese involvement. Cray also buys some high-performance magnetic disks from Fujitsu, as well as units from Ibis.

ETA Systems - The ALSI 20K circuits in the ETA-10 are sourced from domestic vendors with current production from Honeywell. A second U.S. source has been signed and is undergoing evaluation. The memory chips are procured from both U.S. and Japanese suppliers. All ECAD used for design of chips and boards was developed internally by CDC/ETA. The board manufacturing technology was developed by ETA. ETA buys high-performance disks from Ibis and plans to offer comparable high performance disk drives manufactured by CDC when they become available.

2. SUPERCOMPUTER TECHNOLOGY DEVELOPMENTS

New Supercomputers and Product Cycles

The supercomputers developed during the past few years are a sharp departure from their predecessors in several important ways. In addition to the expected improvements in basic speed, there are now architectural changes. One of these is the use of multiple CPUs within the same system to provide additional power beyond that available from evolutionary improvements. The inclusion of massive auxiliary memories is a second change. The combination of these two changes has led to broad performance ranges of supercomputer systems in contrast to the fixed configurations of the recent past.

The production cycle for a supercomputer is about five years, although it remains in productive customer usage for twice that period. During the product cycle, enhancements are introduced as newer technology becomes available. This results in improved models with higher performance and larger memories. Of equal interest is the design cycle which lasts three to four years, and which carries a high dependency on the successful development of underlying technologies, all of which rely upon product use, i.e., sales.

National Security Versus Academic and Commercial Applications

Since 1980 more supercomputers have been sold for academic and commercial/industrial applications than for government applications, including those associated with national security. It is expected that this trend will intensify as more commercial/industrial applications are developed and as academic research finds new uses for supercomputers. The availability of systems from non-U.S. suppliers (Japan) will accelerate the spread of supercomputer usage elsewhere in the world.

Emergent/"Infant" Industry

Until about 1980, supercomputers occupied a low volume, specialized niche in the computer marketplace. The supercomputer segment lagged behind the growth of other products in an industry that was undergoing explosive growth. With the emergence of new technologies and software applications' volume production is now a reality -- and a necessity for technological viability in the supercomputer industry. The presence of competition undoubtedly has helped, including the recent endorsement of vector processing by IBM. It is important to note that it was in the comparatively open, competitive environment of the late 1970s and early 1980s that these developments took place. Today, the effects of Japanese government policies and the trade practices that they nurture are distorting conditions in this emerging market, thereby impeding the viability of private enterprise that does not have strong government support.

New Production Methods

The ETA-10 utilizes new manufacturing techniques that are sharp departures from past traditions in two significant ways. The first is the use of a very dense logic chip using CMOS semiconductor technology that would have been deemed too slow a few years ago. The second is the use of a single large board to contain the entire computer. These two advances will reduce the cost of the system and the time required for manufacturing throughput. Substantial investment in these innovative production methods and new technologies are necessary costs of remaining competitive in supercomputer manufacturing.

3. COMPETITIVE CONDITIONS IN THE U.S. MARKET

National Security Versus Commercial/Academic Trends

There is a growing awareness within the U.S. government that supercomputers are a strategic resource and critical to national security. This leads to concerns within parts of the U.S. government that access should be controlled to supercomputers. With the rapid growth in use here and

abroad, there is pressure from academic and industrial users to expand usage. In addition, the availability of the requisite technologies in foreign countries makes access controls a tenuous proposition. The recent DoD attempts to negotiate access controls over the systems to be installed with NSF funding highlight the difficulty of the situation.

Factors Driving Consumption And Use

Price. Supercomputers, like the rest of the computer industry, have seen excellent gains in price performance. The price of the basic systems has remained constant with time, including periods of high inflation, while the computational performance has increased by two orders of magnitude. Standard industry pricing terms and conditions are available including leasing and installment buying. Price is, of course, a critical factor in supercomputer purchases.

Access to Supercomputing Power

User access to the power of supercomputers has been facilitated by the rapidly broadening base within academia and industry, including time sharing services. The wider range of applications is encouraging, and often forces users to turn to the supercomputer for solutions of problems not amenable to other methods.

Software and Compatibility

The current supercomputers tend to adhere to industry standards. This is the case with FORTRAN, the leading scientific programming language. To exploit the wide usage of FORTRAN, the supercomputer vendors offer specialized software to help the user run programs more efficiently and to take advantage of the supercomputer features. In the operating system area, there is a strong movement underway to implement UNIX, a system based on engineering workstations and many smaller scientific computers. As pointed out previously, applications software is being migrated to supercomputers to the point that a vendor cannot be competitive in the marketplace any longer without offering an extensive applications suite. It is in this area that the U.S. presently holds a big advantage over the Japanese. This may be offset by the fact that the Hitachi and Fujitsu systems are based on IBM compatible processors, thereby offering the ability to migrate some applications directly to their systems, although direct migration will not be able to take advantage of the advanced features. Of course America cannot be complacent about its lead in software as the Japanese continue to make great strides in applications development.

Vendor Service and Support

The supercomputer vendors have traditionally supplied training, analyst, software and maintenance support as standard offerings. Systems availability has grown steadily with improvements in the hardware, software and maintenance techniques. Maintenance and services offered by U.S. vendors are equal to or superior to Japanese services.

Buyer/User Concerns

It is difficult to assess the buyer/user concerns about the viability of a U.S. supercomputer industry. The lack of sales in the U.S. is heavily contributed to by poor and uncertain marketing on the part of Amdahl (Fujitsu's surrogate) and the NEC/Honeywell organization which remains a big question mark. The NAS/Hitachi arrangements are unclear. The U.S. consumer has demonstrated a willingness to buy imports (cars and electronics), as has industry and the U.S. Government in purchasing IBM PC and mainframe clones. Indeed, there has been no evidence of U.S. Government willingness to support the U.S. supercomputer vendors during the crucial development stage.

ETA Systems, Incorporated
1450 Energy Park Drive
St. Paul, MN 55108
(612) 642-3400

ETA SYSTEMS

APPENDIX F-3

October 5, 1987

Mr. James F. Decker
Deputy Director
Office of Energy Research
Department of Energy
Washington, D. C. 20585

Dear Mr. Decker:

In response to your request, I wish to submit some recent recommendations of Control Data/ETA Systems to U.S. policy makers on enhancing the competitiveness of our nation's supercomputer industry. You will note that some of these proposals were contained in the White Paper we presented to the Federal Coordinating Committee on Science, Engineering and Technology (FCCSET) in February 1987.

We believe that if the U.S. is to retain its world leadership in the design, development, manufacture and application of supercomputers we must recognize the gravity of the Japanese competitive threat and implement appropriate policy responses.

The recent U.S. government investigation of Japanese trading practices in supercomputers apparently concluded that the government of Japan has mounted an aggressive stragegy to dominate the global market for supercomputers. This strategy includes protecting the home market against access by American vendors; predatory pricing of supercomputers in the U.S. and third countries; and subsidizing technology development in advanced computing. Such activities are increasing Japan's world market share in this critical sector and could repeat the semiconductor scenario with even more ominous consequences for the national security of the U.S.

Control Data/ETA Systems has already set forth detailed recommendations for addressing the question of trade practices in supercomputers in its February 26, 1987 submission to the U.S. Trade Representative. In summary, we recommended an ongoing rigorous dialogue with Japan to achieve a level playing field in supercomputer trade. We urged that this dialogue include the whole range of issues from market access to predatory pricing practices. We did not advocate protective tariffs or government subsidies to prop up the domestic supercomputer industry.

ETA SYSTEMS

We were pleased with the August 1987 exchange of letters between Ambassador Yeutter and Ambassador Matsunaga which appears to open up Japanese government agency procurements through a more transparent process of bidding rules. And we understand that USTR is committed to treating predatory pricing, particularly the so-called "blockbuster" marketing give-aways to universities, as an outstanding unresolved issue in the overall bilateral relationship. Our government must exert pressure to halt predatory marketing practices in U.S. universities.

But a defensive policy alone will not promote the global competitiveness of the American supercomputer industry. Our country requires a proactive, affirmative U.S. government policy of supporting technological excellence in supercomputing. This policy must take at least four forms:

One, the U.S. government should promise system procurements upon successful demonstrations of design goal achievements. A program of guaranteed procurements of supercomputers that satisfy design and performance standards, not unlike that practiced in our fighter aircraft program would be a powerful incentive t technological preeminence.

In this connection, we strongly urge the Department of Defense, NASA and the intelligence agencies to recognize the profound threat of a dependency on imported supercomputer technology and adopt a policy of domestic content in such procurements. Indeed the GATT procurement code contemplates a domestic reservation in national security or defense industrial base procurements. This policy would of course apply to components and peripherals used in these technologically advanced systems.

Two, we recommend that the government establish a formal program of assigning promising supercomputer design proposals to specific government laboratories and agencies which will procure and integrate these systems into their working environments. Such cooperation would naturally focus on software development and applications.

Three, since increasingly supercomputers are product families with a significant range of performance from smallest to largest, we need to relax and simplify export control procedures. Even where a standard set of safeguards is required, the

ETA SYSTEMS

interagency processing of export licenses simply takes too long and does not seem to consider adequately the differing performance ranges. If our government cannot match Japanese license processing time, the American supercomputing industry will not achieve competitive production levels. Improving the licensing process will also demand far improved communications between industry and government officials.

Four, NSF should give broader support to U.S. university procurements of U.S. supercomputers. It is a national disgrace that today there are more supercomputers in Japanese universities than in our own universities, especially considering their later start.

We thank you for still another opportunity to submit our views on government policy for supercomputing and look forward to the final report and recommendations of FCCSET.

Sincerely,

Lloyd M. Thorndyke
President and CEO

c: R. M. Price
L. D. Rice
T. C. Roberts

decker

IBM International Business Machines Corporation

Office of the Vice President
Data Systems Division

44 South Broadway, White Plains, New York 10601

June 16, 1987

Dr. James Decker
United States Department of Energy
ER 1
Washington, D.C. 20185

Subject: LAX Report Update

Dear Dr. Decker:

Enclosed is an IBM statement on our participation in Numerically Intensive Computing. As we discussed on June 4, it focuses on our current efforts in this area and our directions in computing and networking. I would appreciate its inclusion in the update to the LAX Report.

Sincerely,

T. Wladawsky-Berger

/tm
Enclosure

cc: Mr. J. A. Cannavino
Mr. C. E. McKittrick
Dr. Y. Singh
Dr. D. S. Wehrly

IBM In Numerically Intensive Computing

IBM greatly enhanced its product offerings in the Engineering/Scientific - Numerically Intensive Computing arena in 1985 with the announcement of the IBM 3090 family of 370 systems, which provide superior scalar floating point performance, large memories and parallel processing in addition to an integrated Vector Facility (VF). This enabled IBM 370 systems to be applied to high performance, numeric intensive applications. IBM 3090 systems with one or more Vector Facilities are being used by a broad spectrum of customers in industry, government and university environments. In addition to high performance, these systems provide cost-effective vector processing capacity.

IBM 3090/VF systems have been particularly well accepted in universities, where they are used for education and research purposes, addressing the goals of the December, 1982 LAX report. At Cornell University, an IBM 3090 Model 400 with four Vector Facilities is the primary computer in their NSF-sponsored supercomputer facility. This system will be upgraded to an IBM 3090 Model 600E in July. The research at Cornell is primarily focused on mainframe parallelism and large memory exploitation - tools for program development and algorithms to take advantage of parallel systems and very large memories.

The development investment for the IBM NIC product offerings has been augmented by significant skills and resources resident in four Numerically Intensive Computing (NIC) centers located in Kingston, New York, Palo Alto, California, Rome, Italy and Tokyo, Japan. The primary responsibility of these centers is to work with customers and vendors of NIC applications to enable applications to run effectively on the IBM 3090 with Vector Facility. Technical expertise in many disciplines is available at these centers to assist in the development and transformation of applications and algorithms for vector and parallel processing on the IBM 3090. These applications include those in seismic processing, computational chemistry, fluid dynamics and medical imaging, to name a few. The NIC centers also work on systems aspects of numerically intensive processing, particularly graphics/image output, algorithm development and future application research. The NIC Center staffs are supported by the IBM Research Laboratories and Science Centers throughout the world.

In addition, specially trained Systems Engineers work closely with customers to identify and enable codes to run on the IBM 3090 system with Vector Facility. IBM software has been developed and enhanced to assist in this enablement process. These include:

VS Fortran - with automatic and directed optimization for vector and parallel processing; it also includes a significantly improved interactive debugging capability and a "hot spot" analyzer to determine where tuning or restructuring will be most beneficial.

Engineering/Scientific Subroutine Library (ESSL) - with over two hundred engineering/scientific subroutines using vector or scalar code optimized to provide extremely high performance.

Scientific/Engineering Application Director (SCENAD) - provides menu driven application development and production environments.

Operating Systems - the MVS and VM operating systems both support the IBM 3090 Vector Facility and associated tools. Support for a Unix(R)*-like operating system - a requirement for some NIC applications - is being explored.

With the explosion of NIC applications in many environments, the requirements for NIC capabilities has expanded. Greater performance, enhanced ease of use, more enabled applications and extensions to the System/370 architecture and product line are needed to satisfy the growing needs of customers with NIC applications.

More specifically, the following requirements are being addressed addressed in our development efforts:

Very High Performance: Continued growth in available capacity is required to support larger and larger applications and to improve user turn around times. Higher performance will come from faster scalar and vector processing, and from higher degrees of parallel processing.

Processor storage: Gigabytes of storage are required for the efficient execution of leading edge applications.

I/O performance: Transfer rates between channels and I/O devices and between processors must be improved to speeds far beyond those available today in order to strike the proper balance between processor performance, processor storage and I/O.

Tools: More automatic tools and tools featuring simplicity of use are required to enhance the productivity of the user of NIC systems

---

* Unix(R) is a registered trademark of AT&T.

Systems approach: The integration of hardware, systems software and application software to provide solutions to customers' problems is critical. Heterogeneous environments consisting of networks, processors and software exist throughout the industry. These environments need to be facilitated and accommodated by suppliers.

IBM is addressing each one of these areas. Some of our work has become evident in our systems, telecommunications and NIC announcements made since the 1985 Vector Facility announcement. Our close association with National Laboratories and universities - especially the Cornell Center - provide valuable input and stimuli to our development community. This unique partnership with Cornell was fostered by the 1982 LAX report. In the future it will benefit our customers as we use the knowledge gained in our development activities.

Two major directions are fundamental to our strategy:

1. Parallel Processing: While individual processor performance will continue to increase, continued development and support of parallel systems will enable significant advancements in levels of performance. Considerable investment in software, hardware and research is necessary and is being made.

2. Computer Networking: Communications among computers as well as closer integration of workstations and departmental systems are important parts of the IBM approach to our customers' requirements. Improvement and expansion of IBM networks and protocols and accommodation of non-IBM networks and protocols can be expected. It is IBM's intent to make such transitions transparent to the end-user.

IBM's architectures, systems and products will continue to evolve to address our customers' needs. Today, IBM is supporting considerable research both inside and outside of IBM on technologies, machine structures, and architectures which will bear fruit in the future for the numerically intensive compute environment. IBM recognizes that this is a growing area of opportunity that justifies the required investment and commitment.

June 16, 1987

QUICK, FINAN & ASSOCIATES, INC.
SUITE 340
1020 NINETEENTH STREET, N.W.
WASHINGTON, DC 20036

TELEPHONE (202) 223-4044
TELECOPIER (202) 296-0085

May 19, 1987

Dr. James F. Decker
Deputy Director
Office of Energy Research
Chairman of the Federal Coordinating
Council for Science, Engineering,
and Technology (FCCSET) Committee
on High Performance Computing
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585

Dear Mr. Decker:

I am pleased to have the opportunity to present the views of the U.S. semiconductor industry on an important issue of concern to the Federal Coordinating Council for Science, Engineering and Technology (FCCSET)--the role of the U.S. semiconductor industry in maintaining U.S. leadership in supercomputer technology. Let me briefly describe the current state of conditions in the semiconductor industry, review the special conditions present in the market segments that underpin the supercomputer industry, and lastly, suggest a course for investigating a possible way to deal with some of the concerns over U.S. capabilities to supply key supercomputer-related devices.

A. Current Conditions in the U.S. Semiconductor Industry

The U.S. semiconductor industry today is emerging from a prolonged period of contraction. The following table summarizes what has happened over the past three years to the merchant industry's revenues, earnings, net operating margin, and employment.

	1984	1985	1986
- Worldwide U.S. Revenues ($Bil)	$14.0	$10.8	$11.4
- Pre-Tax Income ($Bil)	$ 2.1	($ 0.8)	($ 0.5)
- Net Operating Income/Sales (Percent)	31	20	23
- Employment (Thou)	473	433	414

Of the total pre-tax losses of $1.3 billion in 1985 and 1986, over half came in the memory markets where Japanese dumping was evidenced. One reason why the U.S. semiconductor industry sought to obtain a government-to-government settlement of the dumping cases on EPROMs and DRAMs and effective implementation of that settlement was to prevent a recurrence of that kind of financial bloodletting in the future. We believe the settlement was designed in a way that seeks to protect U.S. chip consumers from adverse consequences.

Despite the contraction of the market and the major financial damage done to the industry by the Japanese actions--both reflected in the sharp narrowing of net operating margins in the industry--the U.S. semiconductor industry sustained its commitment to investing in R&D. Spending on R&D was $1.3 billion in 1984, $1.5 billion in 1985, and $1.6 billion in 1986, or nearly 14 percent relative to sales revenues. This fact should dispel any idea that we are not committed to the longer-term health of our industry and our customers who benefit from our technology. Despite this commitment to R&D, for structural reasons explained below, it has been difficult for the U.S. semiconductor industry to fully support the specialized needs of the U.S. supercomputer makers.

B. Special Conditions in the Device Markets Related to Supercomputers

Supercomputers are made possible to a large extent by semiconductor technology. But the scale of the supercomputer market from the standpoint of device manufacturers is fairly small--well under one percent of total IC demand and less than what one would roughly calculate would be the minimum efficient scale for a single device manufacturer. A few numbers will make the point:

- In 1986, the high-volume generic integrated circuit (IC) markets accounted for about 16 billion in units worldwide. Unit volumes for some of the basic product groups would be 1.5 billion for memory or 3 billion for MOS logic.

  A slightly different way to look at market scale is to look at a rough estimate of the unit volume per basic product family for major technologies. These figures would be:

for DRAMs	100 million per product family
for SRAMs	15 million per product family
for Bipolar logic	15 million per product family
for MOS logic	15 million per product family

To an individual firm, these figures suggest that an annual production run of roughly 10 million units in DRAMs or 1 to 3 million units in the SRAM or logic area is a reasonable first order approximation for the efficient scale of operation for a firm.

- The total device requirements for supercomputers today is probably on the order of magnitude of 10 million annually; approximately 3 million units would be for various types of memory and 7 million units would be for different types of logic--perhaps under 1 million logic devices being required for especially high-speed cache memory operations. These figures suggest that the scale of the market for some of the highly specialized devices used in supercomputers can only support a very limited number of device manufacturers.

In addition to the problem of limited market size, supercomputers also have especially stringent device requirements with an emphasis on high-speed switching or minimum access times. The leading edge ECL technology used in supercomputers, for example, can have gate delays on the order of 100 picoseconds--roughly 20 times faster than current leading edge CMOS technology--in the region below 10,000 logic gates. These special performance characteristics means that incremental R&D resources--relative to the market opportunity--are necessary to support these device technologies.

The requirement that some devices used in supercomputers have highly demanding, specialized performance parameters, along with the limited market scope, makes it a very challenging market segment to support. Given these conditions, differences between the structures of the U.S. and the Japanese semiconductor and the U.S. and the Japanese supercomputer industries become a factor. The structural conditions in the Japanese semiconductor and supercomputer industries creates a different set of economic incentives relative to those found in the corresponding U.S. industries. Because the Japanese supercomputer firms are part of vertically integrated firms that include semiconductor operations, the Japanese supercomputer designers can call upon their internal semiconductor operations to produce devices for their special applications.

On the other hand, U.S. semiconductor merchant firms must try and earn an adequate rate of return on their limited technical development resources across all available market segments. The U.S. supercomputer manufacturers must compete to attract sufficient resources of the semiconductor merchants to develop devices especially critical to overall system performance.

As a result of these market considerations, to the best of my knowledge, U.S. firms supply only a limited amount of the specialized memory devices used in supercomputers. The story for logic devices is different. In that segment, U.S. firms do provide a substantial portion of the logic devices used in supercomputers--perhaps as much as 90 percent. The small portion of the logic requirements not serviced by U.S. firms is in the area of the very high-performance logic devices where the market potential is most limited. Japanese firms are driven by their internal requirements to supply those specialized devices.

C. How to Address the Structural Issue?

The U.S. semiconductor industry is aware of the problem facing the U.S. supercomputer industry. Clearly, the structural issue outlined above must somehow be addressed. There are several alternative possibilities that can be considered. The most promising one involves an industry initiative just taking final shape. As you probably know, the U.S. industry has undertaken to establish a consortium called SEMATECH, which is targeted on developing future generation processes and equipment

ahead of the product cycle. Current plans envision a prototype facility at SEMATECH, with substantial work performed by subcontractors--usually equipment and materials suppliers--and strong relationships with government-sponsored agencies and laboratories, as well as universities and research agencies. SEMATECH's goals include driving toward technologies that allow companies to perform high-volume, cost-effective manufacturing at one-half micron geometries by the next decade. There will be an emphasis on automation and production flexibility, and the technology will be transferable as individual modules or as a total system. SEMATECH will substantially strengthen the chip production infrastructure on which the entire semiconductor industry--both high and low volume--rely. Current plans call for a chief executive officer of SEMATECH to be hired by September.

I hope this information is of use to the FCCSET Committee, and I encourage you to explore the matter further with the Semiconductor Industry Association and the SEMATECH organization as it develops.

Sincerely,

William F. Finan

William F. Finan
Consultant to the
Semiconductor Industry
Association

MITI INVOLVEMENT IN INFORMATION-RELATED TECHNOLOGY DEVELOPMENT

Tokyo KIKAI SHINKO in Japanese Jan 87 pp 31-36

[Article by Kazuyuki Motohashi of Electronics Policy Department, Machinery & Information Industries Bureau, Ministry of International Trade & Industry (MITI)]

[Text] Supported by technological breakthroughs in recent years, the informationalization of Japanese society is proceeding smoothly. And the information industry that supports this information society continues to exhibit rapid growth as it develops into a leading industry that is pioneering the way to the 21'st century.

To be sure, informationalization not only makes industry and society more active and sophisticated, but also contributes enormously in making everyday life less onerous and more convenient, and in promoting mutual understanding in the international community. The capital investments which accompany informationalization, moreover, provide a stable stimulation to domestic demand. This is also very beneficial to the Japanese economy.

However, informationalization in Japan has so far been limited primarily to individual industrial fields, and many tasks remain to be done before high-level informationalization can be achieved, such as creating inter-industrial systems and deploying omnifunctional information networks.

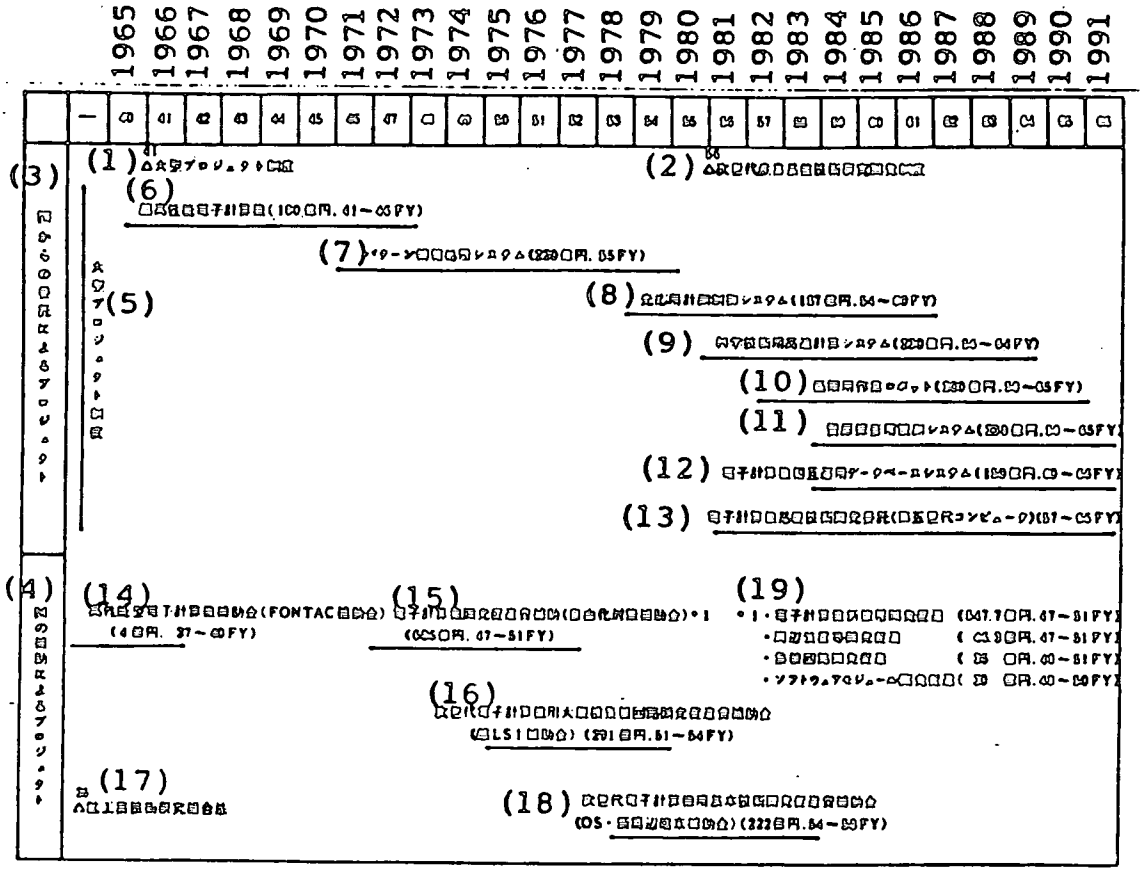
In this article I wish to discuss MITI's involvement in developing information-related technology, which is crucial to successfully tackling the tasks which face us in implementing greater informationalization.

Development of Policies on Information-Related Technology

MITI has shown awareness of the importance, far-reaching influence, and future potential of the computer industry--as an advanced industry--since very early on, and has implemented various policies in that regard, such as the Industrial Testing Grant in 1950 (shortly after the war), and the Mining Technology Research Grant (cf Figure 1).

To mention some of the more important measures, there is the "FONTAC Grant" system (1962-) aimed at the development of large domestic computers, and

Figure 1 Projects Related to Development of Information Technology



Key:

1. (1966) ^ Large-Scale Project Program
2. (1981) ^ Next Generation System
3. Projects commissioned by national government
4. Projects subsidized by national government
5. Large-Scale Project Program
6. Ultra-High-Performance Computer (10 billion yen, FY 1966-1971)
7. Pattern Data Processing System (22 billion yen, FY 1980)
8. Optical Measurement Control System (15.7 billion yen, FY 1979-1985)
9. High-Speed Computer System for Science & Technology (23 billion yen, FY 1981-1989)
10. Robots for Extreme Operations (20 billion yen, FY 1983-1990)
11. Resource Exploration Monitoring System (23 billion yen, FY 1984-1990)
12. Computer-Interoperable Database System (15 billion yen FY 1985-1991)
13. Commissioned Development of Basic Computer Technology (fifth-generation computer) (FY 1982-1991)
14. High-Performance Computer Subsidy (FONTAC Grant) (400 □illion yen, FY 1962-1965)

15. Computer Development Promotion-Expense Subsidy Program (Liberalization Policy Subsidy)*1 (68.6 billion yen, FY 1972-1976)
16. Subsidies for Promotional Expenses for Development of VLSI's for Next-Generation Computers (VLSI Grant) (29.1 billion yen, FY 1976-1979)
17. (1961) ^ Mining Technology Research Cooperative Act
18. Subsidies for Promotional Expenses for Development of Basic Technology for Next-Generation Computers (OS/New Peripherals & Terminals Grant) (22.2 billion yen, FY 1979-1983)
19. *1 — New computer model development promotion (54.77 billion yen, FY 1972-1976)
    — Peripheral equipment development promotion (4.63 billion yen, FY 1972-1976)
    — Integrated circuit development promotion (3.5 billion yen, FY 1973-1976)
    — Software module development promotion (3 billion yen, FY 1973-1975)

the "Computer Development Promotion-Expense Subsidy Program" that accompanied the "Computer-Related Liberalization Measures" of 1971 and which was intended to fundamentally strengthen the domestic manufacturers and enhance Japan's ability to develop computer technology. Besides such subsidy programs as these to aid industry, MITI has also implemented the "Large-Scale Industrial Technology Research & Development Program" (called the "Big Project"), which was a commissioning project meant to bring more technological know-how to Japan.

Until the 1980's, MITI's R&D projects had been informed by a "catch up with the West" mentality. Then came a shift away from this approach, as MITI began to emphasize the necessity for a more aggressive approach in the international community involving "seed-planting & nurturing" technological development. In recognition of this necessity, priority is now given to national projects on the world's technological frontiers, such as the "R&D System for Next Generation Industries" (the so-called "Next Generation System"), and fifth-generation computer R&D.

Position of Information-Related Technology Policies

As a result of these various subsidy programs for aiding the private computer industries, and the private sector's aggressive approach to information technology, the communications, electronics, and electric measuring instrument fields have grown even faster than other fields in terms of R&D expenditures, which have been growing steadily accross the board in Japan (cf Figure 2).

However, since this R&D involves enormous risks, prolonged gestation, and very large capital outlays, one could not expect it to proceed steadily if it were left entirely up to the private sector. Thus it is necessary for the national government itself to undertake research and development in fields that are especially critical to the national economy.

Figure 2 Research Expenditure Trends for Main Manufacturing Industries

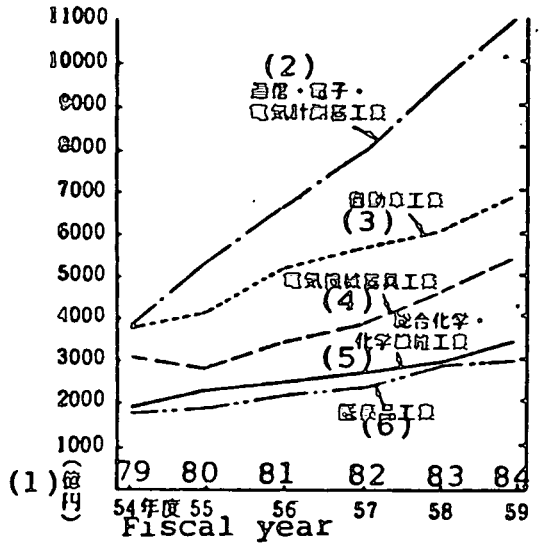
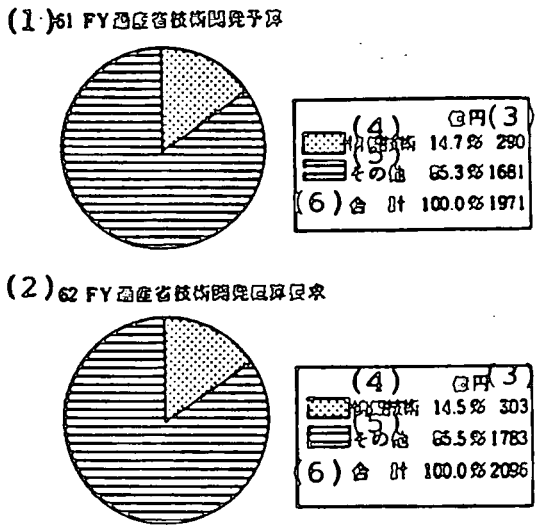


Figure 3 Information Technology Share in MITI Technology Development Budget



Key to Figure 2:
1. (Units = 100 billion yen)
2. Communications, electronics, and electric measurement instruments
3. Automobile industry
4. Electric machine & tool industry
5. Gen chemicals, chemical textiles
6. Pharmaceutical industry

Key to Figure 3:
1. MITI technological development budget for FY 1986
2. MITI technological development estimated budget for FY 1987
3. Units = 100 billion yen
4. [dots] Information technology
5. [lines] Other
6. Total

Out of such considerations as these, the national government sponsors R&D projects by commissioning research work to private industry, in "areas having wide-ranging influence on the economy, society, and technology, but requiring a long time to implement," "fields in which the size of the risks involved and the capital required for development surpass the capacities of private industry," "fields in which the needs of the economy or the society are extremely great and pressing, requiring immediate action," and "fields in which there are public needs that require various kinds of social and institutional coordination," etc. Examples of such projects are the Fifth-Generation Computer Project, the "Big Project," and the Next Generation System. In addition, the national government itself undertakes research projects in the area of basic research which will give rise to creative and autonomous technologies in the future. Such work is done at places like the Electrotechnical Laboratory.

Information technology accounts for roughly 15 percent (about 25 billion yen) of both the 1986 budget and the 1987 estimated budget for technology

development programs overall (cf Figure 3), and this percentage is expected to increase in the future.

MITI Programs for Information Technology Development

As discussed above, MITI is keenly aware of the importance of information technology and of the need for government-sponsored projects, and the ministry has implemented various measures in the interest of developing such technology (cf. Figure 4). We may divide such technological development into three categories, namely the privately sponsored projects supported by Key Technology Center funding, the national R&D projects funded directly by MITI, and the research carried on by the Electrotechnical Laboratory (which is a national research institution). A broad range of technical fields are covered, including materials & devices, data processing, communications, and space.

Figure 3 MITI Programs for Information Technology Development

Promotion of Technology Development (business scale) FY86: 22.4 billion yen, FY87: 22 billion yen (govt draft proposal) (excluding Key Technology Center and Electrotechnical Laboratory funding)

Materials, Device Technology [U = undetermined, * = repeat item]

FY86: (1) 1542, (2) 2889, (3) 635, (4) 750 (Units: million yen)
FY87: (1) 1441, (2) 2946, (3) U , (4) U

- (1) Development of new function elements (Next Generation System, FY81-86, 7.8 billion yen) —> 3D circuit elements, superlattice elements, bio-elements
- (2) High-speed computer system for science/technology (Big Project, FY81-89, 23 billion yen)—> super-high-speed element
- (3) Optical-electronic IC R&D (Key Tech Ctr funding, FY85-95, 100 billion yen) —> OEIC
- (4) Synchrotron orbital radiation (SOR) (FY85-95, 14.3 billion yen) —> SOR (X-ray lithography)

Information Processing Technology

FY86: (1) 5491, (2) 1260, (4) 830, (5) 30, (6) 600, (7) 2800
FY87: (1) 5631, (2) 1355, (4) 1055, (5) 167, (6) 3000, (7) 3000

- (1) Fifth-generation computer project (FY82-91, approx 20 billion yen up to FY86) —> AI, natural language processing, machine translation, machine interface

- (2) Special development of software technology (development commissioned by IPA, FY82-) —> Software environment integration technology, etc.
- (3)° High-speed computer system for science/technology (Big Project, FY81-89, 23 billion yen) —> High-speed operations, large-capacity high-speed memory, distributed processing machines, etc.
- (4) Computer interoperable database system (Big Project, FY85-91, 15 billion yen) —> multimedia, distributed database, etc.
- (5) Development of machine language systems for use between neighboring nations (FY86-92, 6.25 billion yen) —> machine translation
- (6) Development of electronic dictionaries for natural language processing (Key Tech Ctr funding, FY85-94, minimum 14.3 billion yen) —> Language-concept knowledge base for fifth-generation computers
- (7) Development of system for industrializing software production (FY85-89, 25 billion yen) —> Sigma project

Communications Technology

FY86: (2) 370 FY87: (2) U

- (1)° Computer interoperable database system (Big Project, FY85-91, 15 billion yen) —> Development of data transmission software, OSI promotion, establishment of interoperability conformity
- (2) Basic measurement technology for coherent optical communications (Key Technology Ctr funding, FY85-91, 4.3 billion yen) —> Laser technology, high-efficiency high-density high modulation

Space Technology

FY86: (1) 4391, (2) 475 FY87: (1) 3142, (2) U

- (1) Resource exploration monitoring system (Big Project, FY84-90, 23 billion yen) —> Composite aperture radar, engineering sensors, high-speed high-volume transfer technology

- (2) Comprehensive research on space observation applications (Key Technology Ctr funding, FY85-92, 5.7 billion yen) —> Development of systems for space environment experiments

Applications Systems

FY86: (1) 103, (2) 2405, (3) 520, (4) 682 (Units: million yen)
FY87: (1) 169, (2) 2425, (3) U , (4) 669

- (1) Diagnostic support system (FY82-88, total amt undetermined) —> MEDIS
- (2) Extreme operations robots (Big Project, FY83-90, 20 billion yen) —> Image recognition, etc.
- (3) High-level information-processing video information system (Key Technology Center funding, FY85-90, 4.8 billion yen) —> Advanced HI-OVIS
- (4) R&D commissioning program for medical & welfare equipment technology (22 devices, FY76-) —> CT scanner, nerve damage diagnostics, treatment support systems

Electrotechnical Laboratory

- (1) Materials Dept (electronic materials, magnetic & amorphous materials, etc.)
- (2) Electronic Device Dept (high-speed elements, new function elements, etc.)
- (3) Pattern Information Dept (Voice & image recognition & comprehension technology bionics, etc.)
- (4) Computer Dept (data processing, storage, I/O technology)
- (5) Software Dept (programming languages, network architecture)
- (6) Control Dept (information systems control technology)
- (7) Radio Electronics Dept (laser, optoelectronics technology)
- (8) Extreme Technology Dept (space environment technology, etc.)

We will now introduce a few of the national projects which we believe to be particularly important among the various programs for technological development.

(1) Fifth-Generation Computer Project

Computer technology has recently moved into the fourth generation, characterized by VLSI implementation. With the rapid progress being made in computer technology, together with the remarkable changes occurring in the

economic and social environments, it is believed that the new generation of the 1990's will demand radically new computer architectures that overcome many of the shortcomings of conventional computers.

Against this background, MITI began conducting preliminary surveys in fiscal 1979, and started R&D work on the fifth-generation computer in fiscal 1982.

The entire 10-year project is divided into an initial phase (3 years), middle phase (4 years), and final phase (3 years), taking a step-by-step approach to the development of a revolutionary new computer by fiscal 1991 that will function as a highly sophisticated AI system.

This year marks the second year of the middle phase, or the fifth year of the overall project. With the development of the sequential inference machine (SPI, marketed by Mitsubishi Electric Corporation) during the initial phase, and other successes, the project is moving along on schedule.

The framework of this R&D is such that the Institute of New Generation Computer Technology (ICOT) is commissioned by the government. The fifth-generation computer project is being warmly praised among the advanced nations as well as in Japan. It is no exaggeration to say that the eyes of the whole world have been focused on ICOT. The U.S. National Science Foundation (NSF) requested in June of last year that they be permitted to send long-term researchers to participate in ICOT, and memoranda have been exchanged, resulting in the request being granted.

In any event, this project represents a most critical technological development program for establishing Japan on the frontier of world technology, and MITI has plans for pushing the project ahead even more aggressively.

(2) Computer Interoperable Database System

With the proliferation of mass communications and the advance of computerization, the structure of the information society in Japan has become increasingly advanced in recent years. Human living and activity in the society of the 1990's will require vastly greater volumes of information than they do today, and people must be able to handle this information, as well as to generate large quantities of various kinds of information themselves. In the midst of this advancing informationalization, it will be necessary to build highly reliable systems with which databases of various information can be jointly used if we are to see the information society develop in Japan the way it should.

However, database technology that can cope with multimedia is just now in the stage of initial R&D. Due to differences among computer makers, not only is the equipment not interoperable, it cannot even be connected.

In order to resolve such problems as these, we need to make computers that can handle such diverse media as text, graphics, video, and audio, to build

diversified databases that will cope with these media and be distributed widely, and then put all of this together in a network so that it can be mutually useful.

From this perspective and based on the Big Project program, the Interoperable Database System project aims to develop the necessary gechnology to develop advanced connecting methods and techniques and to build a network system through which interoperability will be achieved, conducting research and development in various fields to enhance the reliability of multimedia distributed databases and systems.

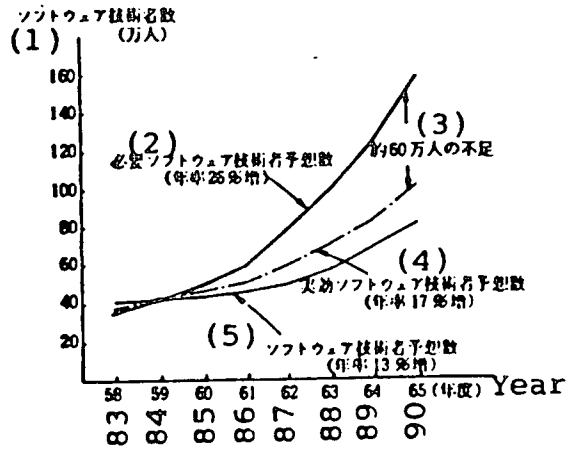
This project is also important for its support--in terms of technological development--of the OSI (open systems interconnection) model being promoted worldwide by the ISO (International Standards Organization).

(3) System for Industrializing Software Production (Sigma System)

Informationalization is advancing rapidly in Japan, and the importance of computer systems to economic and societal activity is rising dramatically. This has led in recent years, however, to an increasingly strong awareness of the large problem posed by the so-called "software crisis."

As informationalization has proceeded rapidly, a software supply-and-demand gap has developed and is widening. This is the essence of the software crisis. Software demand is reportedly expanding at an annual rate of 26 percent. Technological progress in the area of software development, however, is woefully inadequate, with software production growing annually at a mere 4 percent or so (according to W. E. Liddle). Thus software production cannot catch up to the demand. The number of technicians working on software development, moreover, is growing at a rate of only 13 percent annually, and it is projected that by the year 1990 we will face a shortage of 600,000 software specialists (cf Figure 5).

Figure 5 Projected Software Supply-Demand Gap in Japan



Key:
1. Number of software technicians (units = 10,000)
2. Estimated number of software technicians needed (26% annual growth)
3. Approximate shortage of 600,000 technicians
4. Estimated effective number of software technicians (17% annual growth)
5. Estimated number of software technicians (13% annual growth)

In order to overcome this software crisis and promote smooth informationalization in Japan, a number of strong measures must be taken, including (1) raising software-development productivity and reliability, (2) promoting the training and retention of software specialists, and (3) promoting the proliferation of general-purpose software. The first of these, namely the sharp upgrading of software-development productivity and reliability, is the most urgent task now facing us in this area.

Informed by this perspective, MITI began promoting the development of the Sigma System in 1985 as an IPA (Information Processing Promotion Association) project.

The Sigma System is composed of three elements, as follows.

-- Sigma Center (Sigma System control and management; database management)

-- Sigma Network (Logic network linking Sigma Center and users)

-- User Sites (Users' computer systems connected to Sigma Network)

The system seeks to improve software productivity and reliability by means of the above three subsystems.

Future Measures

(1) Promotion of International Cooperation

Japanese-American trade friction has gone through several stages. In the early 1970's it involved textiles and steel. In the late 1970's it was electrical appliances such as color televisions. Now the focus of this trade friction is on such high-tech fields as communications, electronics, and semiconductors. There is a growing awareness in Europe and America, moreover, that the overwhelming competitive power of Japanese merchandise exports is the result of industrial applications of the fruit of basic research done originally in Europe and America. This has given rise to discontent over the so-called technological free-ride, and is provoking much anti-Japanese sentiment, especially in the United States. One countermeasure against this criticism which should be taken--while aggressively publicizing Japanese basic-research projects--is to make it clear how the Japanese are contributing to basic research fields that benefit the whole world.

Informed by this perspective, we believe that international research exchanges, such as the aforesaid memoranda exchanged between ICOT and NSF, and technological aid to developing countries, will become increasingly important.

(2) Utilizing Private Enterprise

The portion of technological development costs now contributed by the Japanese government amounts to about 25 percent, which is lower than in the western nations. Increasing the government's contribution is an urgent necessity, but when one considers Japan's tight fiscal situation in recent years together with the booming prosperity now enjoyed in the private sector, the continuing utilization of private enterprise is going to determine the direction and extent of future technological progress in Japan.

The specially licensed corporation called the Key Technology Center that was formed with financing from private industry in October, 1985, is seeking aggressively to engage private enterprise in basic research, and its activities bear careful watching.

In addition, five companies involved in information technology were formed in fiscal 1985 with Key Center financing, as follows (cf Table 1).

(1) Nippon Electronic Dictionary Research Institute, Ltd (Bun'ichi Oguchi, president)
(2) Sortech, Ltd (Tsuneo Shio, president)
(3) Optical Technology Research & Development, Ltd (Noriyuki Uenohara, president)
(4) Optical Measurement Technology Research & Development, Ltd (Shozo Yokogawa, president)
(5) Key Information Systems Development, Ltd (Takayoshi Shirozaka, president)

In the foregoing we have discussed the global situation with regard to information technology development, and Japan's activities in that field. In the future, we believe that Japan will play an increasingly major role in actively expanding the world's technological frontiers in basic research fields, as is symbolized by the fifth-generation computer project. This activity will be focused on private enterprise, in a wider context of international cooperation.

Table 1 Information-Technology-Related Key Technology Center Financing for Fiscal 1985

Project Name: Electronic Dictionary for Natural Language Processing
Company Name: Nippon Electronic Dictionary Research Institute, Ltd
Fiscal 1985 Key Center Financing: 200 million yen
Financers (scheduled): Fujitsu Ltd, Toshiba Corp, Hitachi Ltd, Oki Electric Industry Co, Ltd, and four others

Basic Plan: Will attempt to draft a prototype large-scale electronic dictionary that covers a wide range of areas, such as is required for programming computers to recognize the language used ordinarily by humans (natural language processing). This will require the comprehensive and systematic collating of an enormous quantity of linguistic data, and the rendering of this data into machine-readable form.

Project Name: Optical-Electronic IC Research & Development
Company Name: Optical Technology Research & Development, Ltd
Fiscal 1985 Key Center Financing: 100 million yen
Financers (scheduled): Nippon Electric Co, Ltd, Oki Electric Industry Co, Ltd, Sumitomo Electric Industries, Ltd, Toshiba Corp, and nine others
Basic Plan: An optical-electronic integrated circuit (OEIC) is an integrated circuit that simultaneously implements photoelements and electronic circuitry on a single crystal substrate. Transmission speeds of 1 gigabit/second have been achieved to date. Process and device technologies will be developed with the goal of achieving 10 gigabits/second.

Project Name: Research on Basic Measurement Technology for Coherent Optical Communications
Company Name: Optical Measurement Technology Research & Development, Ltd
Fiscal 1985 Key Center Financing: 90 million yen
Financers (scheduled): Yokogawa Hokushin Electric Co, Ltd, Advantest, Ando Electric Co, Ltd, Iwatsu Electric Co, Ltd, Anritsu
Basic Plan: In order to achieve higher capacities in optical communications, signal modes in which light frequencies instead of light intensities are modulated are being considered. This requires using a light medium in which wave coherence has been enhanced. This project aims to develop practical optical measurement technologies for measuring the attenuation, power, frequency, and phase of such coherent light.

Project Name: Synchrotron Orbital Radiation Technology R&D
Company Name: Sortech, Ltd
Fiscal 1985 Key Center Financing: 150 million yen
Financers (scheduled): Mitsubishi Electric Corp, Toshiba Corp, Nippon Electric Co, Ltd, Hitachi Ltd, and nine others
Basic Plan: When matter is irradiated with light, changes occur in its chemical bonds and structure. This phenomenon is conspicuous in the wavelength region that extends from the vacuum ultraviolet region to the soft X-ray revion (10^{-3} - 10^{-5} cm). In this region of the spectrum, synchrotron orbital radiation (SOR) constitutes a parallel light source that is stronger than other light sources by a factor of 10^2. Industrial SOR applications will be developed in such diverse fields as ultra-micromachining techniques and photochemical reactions.

Project Name: Research on Advanced Video Information Processing Systems
Company Name: Key Information Systems Development, Ltd
Fiscal 1985 Key Center Financing: 80 million yen

Financers (scheduled): Sumitomo Electric Industries, Ltd, Fujitsu Ltd, Matsushita Electric Industrial Co, Ltd

Basic Plan: In order to perfect totally new information systems which have functions for comprehensively processing and providing multimedia information (video, audio, text, etc.), the necessary system design techniques, data processing technology, and component devices will be developed, and R&D will be conducted in the areas of fabricating and experimental systems, and testing their functions and performance.

12312
CSO: 4306/5026

Session 3327

The Impact of Supercomputers on Experimentation: A View from a National Laboratory

Victor L. Peterson
James O. Arnold
NASA Ames Research Center
Moffett Field, California

Abstract

The relative roles of large-scale scientific computers and physical experiments in several science and engineering disciplines are discussed. Increasing dependence on computers is shown to be motivated both by the rapid growth in computer speed and memory, which permits accurate numerical simulation of complex physical phenomena, and by the rapid reduction in the cost of performing a calculation, which makes computation an increasingly attractive complement to experimentation. Computer speed and memory requirements are presented for selected areas of such disciplines as fluid dynamics, aerodynamics, aerothermodynamics, chemistry, atmospheric sciences, astronomy, and astrophysics, together with some examples of the complementary nature of computation and experiment. Finally, the impact of the emerging role of computers in the technical disciplines is discussed in terms of both the requirements for experimentation and the attainment of previously inaccessible information on physical processes.

Introduction

Computers are playing an increasingly important role in the science and engineering disciplines. They are, in fact, revolutionizing not only the way research is conducted in nearly all fields, but also the way that scientific knowledge and wisdom are applied in the industrial environment, including the search for and recovery of natural resources, the design and manufacture of products, the development of pharmaceuticals, and the production of motion pictures. This revolution, still in its infancy, began about 20 years ago. It is gathering momentum with exponential growth, and it eventually could rival other great events in the course of the evolution of the civilized world. It is crucial for those involved in the pursuit of technical endeavors, including education, to understand how the computer revolution is enhancing the economic and technical value of human resources, and to take the necessary steps to stay at its leading edge. The consequences of ignoring this revolution are not acceptable for a country dedicated to commanding a leadership position in world affairs.

The principal objective of this paper is to provide a brief overview of how computers are beginning to affect the technical disciplines and to discuss how they are changing the requirements for associated physical experiments. This subject will be treated by examining some examples related to aerodynamics, aerothermodynamics, chemistry, atmospheric sciences, astronomy, and astrophysics.

Background

The computer revolution in science and engineering is intimately connected to the development of both computers and numerical methods. Therefore, it is appropriate to review past advances and to quantify future prospects. This review will be limited to top-of-the-line computing engines, commonly referred to as supercomputers. These are the machines that are pacing the development of the computational disciplines, although to be used effectively they must be augmented by appropriate peripheral devices such as front-end computers, terminals, graphics devices, long-term data storage facilities, and communications networks. It is also necessary to have skilled persons trained in the use of supercomputers. There is a shortage of such personnel, and there are only a few universities equipped with supercomputers to provide this training. This must be remedied soon if the growing demand for these specialists is to be met.

The increase in computer speed and cost is shown for some existing and planned machines in figure 1.[1] It is noteworthy that computer speed has grown about four orders of magnitude over a period of 30 years, whereas monthly rental cost has risen by approximately a factor of only 10 in actual-year dollars. In terms of the real value of money, the cost has actually decreased substantially. The exponential growth in computer speed is expected to continue for some time as a result of advances in very-large-scale electronics integration and computer architecture technologies. The increase of computer memory, which is shown in figure 2, has been only about half as large as that for computer speed.[1] The rate of growth is

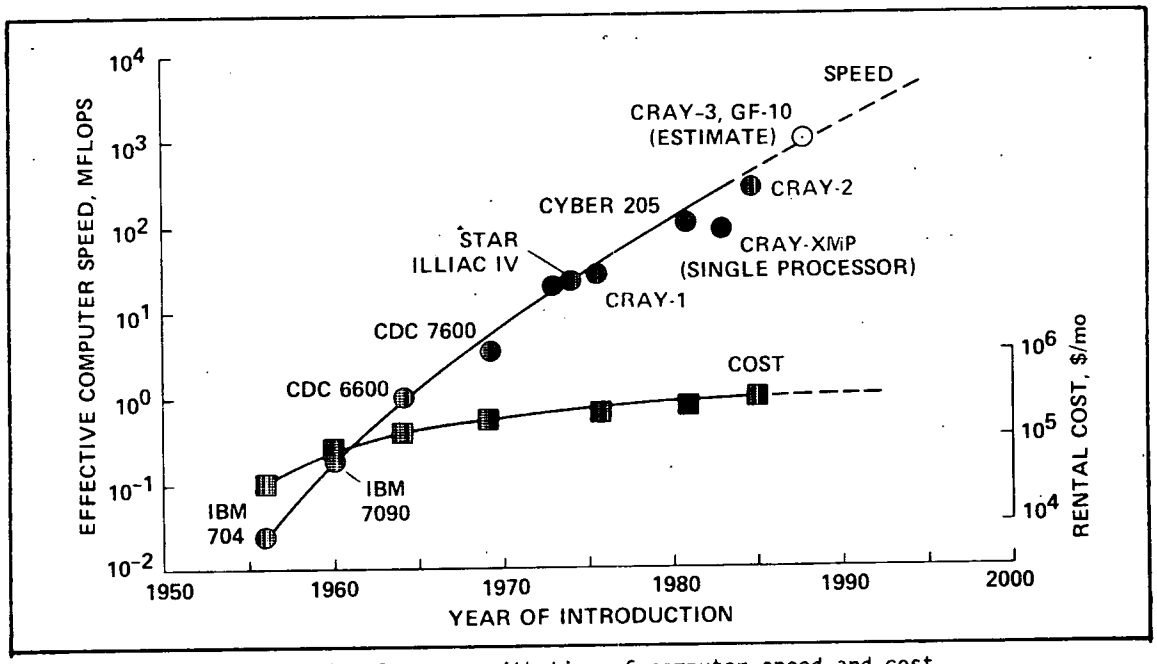


Figure 1.- Increase with time of computer speed and cost.

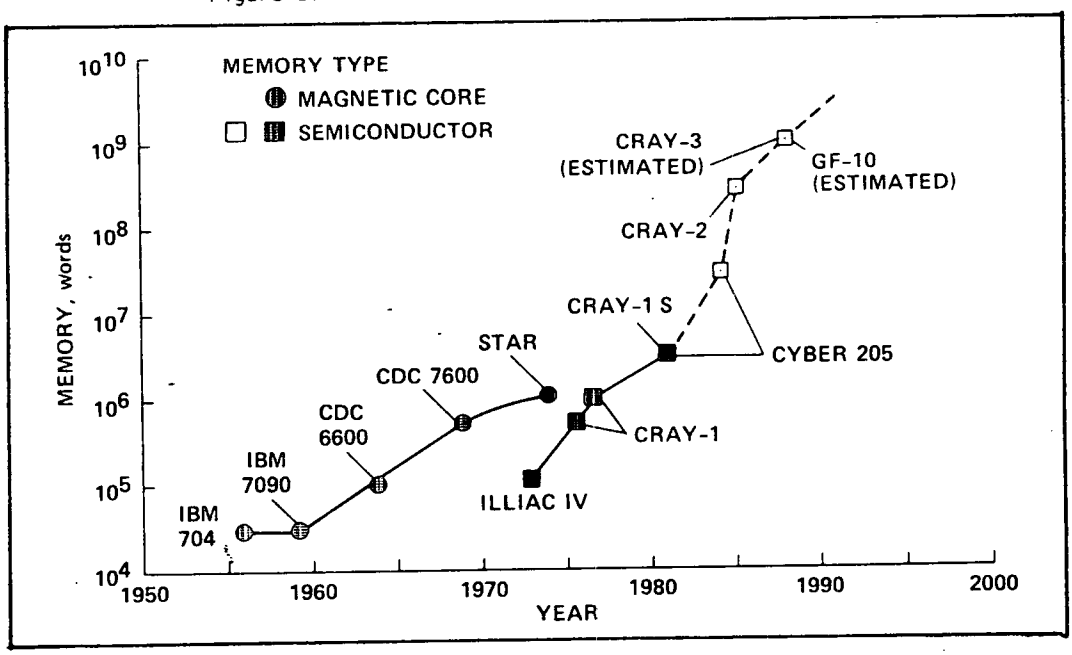


Figure 2.- Increase with time of computer main memory.

expected to be high in the foreseeable future, however. In fact, a Cray-2 computer with a main memory of 256 million 64-bit words is expected to be delivered to the NASA Ames Research Center in the fall of 1985. This capability represents a factor of 8 increase over the 32-million-word Cyber 205 just now being made available. It is almost certain that memory sizes as large as one billion words will be available before 1990.

Improvements in computer performance have been closely paralleled by improvements in numerical methods over the past 20 years. This is illustrated by the data presented in figure 3, which show how the cost of performing a computation has been driven down by the advances that are being made in computers and in numerical methods.[2] In this case, the methods refer to those used to solve two approximating forms of the Navier-Stokes equations governing perfect-gas fluid dynamics, but the results are indicative of methods used for equations governing many other disciplines. Further improvement (several orders of magnitude) in algorithms for solving the Navier-Stokes equations appears to be possible at

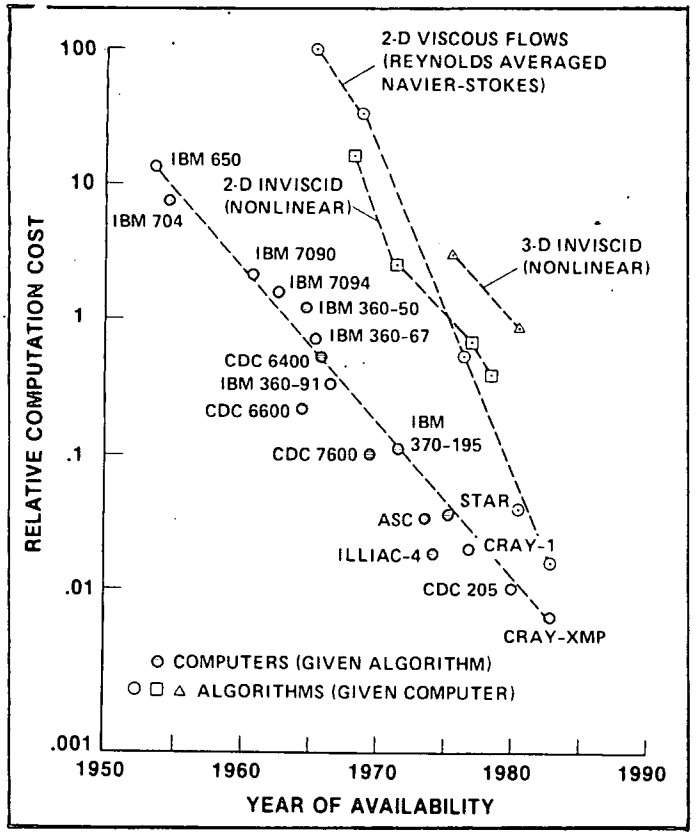


Figure 3.- Comparison of numerical simulatuion cost trend resulting from improvements in algorithms for different forms of the fluid dynamics equations with that owing to improvements in computers.

this time. Results of these improvements in computer and algorithm performance compound to provide a 10^5 reduction, over a 15-year period, in the cost of performing a computation with a given computer code.

Theory and experiment have been used by scientists and engineers in a complementary fashion for many years. In some cases, the physical observation comes first, in other cases, the situation is reversed. Prior to the development of digital computers, however, the role played by theory was hampered by the inability to solve the governing equations for many phenomena. The relative roles of theory and experiment began to change after the introduction of the electronic digital computer. The exact time when these changes begin to occur in a particular discipline is directly related to the complexity of the physics involved and the level of computer power required to solve the governing equations in a reasonable amount of computing time commensurate with the importance of obtaining answers. Thus, different disciplines have different thresholds in terms of computer requirements.

One of the first disciplines to lend itself to practical treatment by the digital computer was the prediction of cannon-shell trajectories. In fact, the ENIAC computer was developed during World War II to calculate ballistics firing tables. Today, problems involving flight mechanics and orbital dynamics are solved routinely without the use of validating experiments. Machines developed in the early 1960s, such as the CDC 6600, were powerful enough to enable research and development of nuclear devices to continue effectively despite the nuclear test-ban treaties. The design of aircraft structures is now done largely using computers, with only minimal proof-testing of the finished product. By 1970 computers were powerful enough to begin tackling nonlinear forms of the equations governing fluid dynamics for simple flow geometries. A little more than 10 years later there were examples of significant geometric design modifications, based solely on computation, being installed on aircraft and committed to flight without the benefit of validating experiments. These are but a few examples of how computers can be adapted for use by various disciplines when they become powerful enough to treat the underlying physical principles with a high degree of precision.

Both technical and economic factors motivate the increasing use of computers in the technical disciplines. While experimental apparatus and advanced instrumentation enhance the ability to observe, computers enhance the ability to reason. Phenomena governed by the laws of physics can be expressed in mathematical form, and computers can be used to solve the resulting equations with a degree of exactness not otherwise possible. In many cases, computers enable researchers to study phenomena that are difficult, if not impossible, to study experimentally. Computers can assist engineers and scientists in meeting increasingly complex demands placed on industrial design and engineering. These demands arise from the need to be competitive on a global scale. There is very little tolerance for design errors despite conflicting requirements of low cost, high quality, energy efficiency, and environmental compatibility. In addition, computers enable industrial product developments to evolve more rapidly to assure economic market success. One by one, the country's major industries are acquiring supercomputers. This expanding market should help control the future costs of large-scale machines.

Unlike many experimental test facilities and instrumentation acquired through discrete purchases and then used for many years without major change, computational equipment requires frequent and continual upgrading to remain current. Supercomputer performance has been increasing by a

factor of 10 about every 7 years. Thus, the mainframes should be replaced long before their components are no longer serviceable (typically 15 years). Peripheral devices are also changing rapidly. For example, over a 20-year period, the method of inputting data into the computer has changed from reading punched cards at the site of the computer, to reading punched cards fed in through remote job-entry stations, to the use of "dumb" terminals on the user's desk, to the use of intelligent work stations in the user's office. Likewise, the method of providing computer output has changed from tabular listings, to centralized plotting, to remote plotting, to distributed color graphics devices. Therefore, a modern computing environment requires an annual budget considerably larger than that required to maintain and operate existing equipment, and it requires a careful and continuous planning process to stay abreast of the rapidly changing technology.

The availability of large computers is also affecting the cost of performing experiments. On one hand, the cost is going down because preliminary work done on the computer can be used to guide the conduct of experiments and often eliminates the need to perform measurements for numerous values of the variables. The computer, in effect, provides a great deal of insight into what to expect from the experiment before it is conducted, and it provides a means for interpolating between widely spaced values of the parameters. On the other hand, the cost is going up because computer-generated information provides much more detail concerning the underlying physics than can be obtained from more analytical examination, and thus inspires new experimental requirements. The experimentalist is being forced to conduct more sophisticated experiments that require the development of improved, but costly, test techniques. Eventually, the costs associated with experiments in selected fields can be driven down, if not eliminated, as computers become sufficiently powerful to treat situations without appreciable approximation.

Impact of Computers on Selected Disciplines

The degree to which computers are affecting the scientific and engineering process exercised in the various disciplines depends, to a large extent, on the complexity of the underlying physics and the degree of exactness with which available computer power can be used to solve the governing equations. Since the amount of required computer power varies from discipline to discipline, each area of interest must be examined individually to determine the impact. The process involves making an estimate of the computer speed and memory required to solve the governing equations, comparing the requirements with computer capabilities, and then factoring in findings based on past experience. Computers are not yet large enough to treat all situations from first principles so the ability to simplify with models or approximations plays a strong role in most disciplines. Results of applying this process to several disciplines are discussed in this section.

Fluid Dynamics

The Navier-Stokes equations govern the motion of viscous fluids over a broad range of continuum-flow situations, including those of interest in aircraft design. It is not possible to obtain closed-form solutions to these equations for practical engineering problems. However, various degrees of approximation have been worked out over the years to obtain useful results.

Four major levels of approximation to the full equations have been identified.[3] Each level of approximation resolves the underlying physics to a different degree, provides a different level of understanding, and requires a different level of computer capability. These approximations, their capabilities in resolving problems associated with aircraft aerodynamics, and the computer requirements to solve them in a reasonable amount of time (about 15 min) are summarized in table 1 and discussed in some depth in the literature.[2,3] Computer requirements are expressed in terms of the power of a Class VI machine which is defined here to have a processing speed of 30 million floating-point operations per second (MFLOPs) and a main memory of about 8 million words. Computer requirements increase with each higher level of approximation, both because more flow variables are involved and because more panels or grid points are required to resolve the flows to a level of detail that is commensurate with the physics embodied in the approximation. Experience indicates that the Reynolds-averaged form of the Navier-Stokes equations probably will be adequate for most design-oriented problems. The effects of all scales of turbulence are modeled in this level of approximation. The development of these turbulence models is the subject of extensive current research by both computational and experimental fluid dynamicists. In fact, the experimentalists are being guided, to a large extent, by computational research programs which are based either on the application of the large-eddy simulation approximation or on the use of the full Navier-Stokes equations for simple flow geometries. Information presented in table 1 shows that computers having one-tenth the power of a Class VI machine are required to begin to make inroads on the computational treatment of aircraft design problems. Machines of this class, such as the CDC 7600, became available in the late 1960s, but it was not until the mid-1970s that they were generally accessible by the aerodynamicists. Now, of course, Class VI machines, such as the Cray-1 and the Cyber 205, are widely available.

Table 1.– Major levels of approximation to the Navier-Stokes equations with results provided, and computer requirements to obtain solutions in 15 min of computation time.

APPROXIMATION	CAPABILITY	GRID POINTS REQUIRED	COMPUTER REQUIREMENT
LINEARIZED INVISCID	SUBSONIC/SUPERSONIC PRESSURE LOADS VORTEX DRAG	3×10^3 PANELS	1/10 CLASS VI
NONLINEAR INVISCID	ABOVE PLUS: TRANSONIC PRESSURE LOADS WAVE DRAG	10^5	CLASS VI
REYNOLDS AVERAGED NAVIER-STOKES	ABOVE PLUS: SEPARATION/REATTACHMENT STALL/BUFFET/FLUTTER TOTAL DRAG	10^7	30 × CLASS VI
LARGE EDDY SIMULATION	ABOVE PLUS: TURBULENCE STRUCTURE AERODYNAMIC NOISE	10^9	3000 × CLASS VI
FULL NAVIER-STOKES	ABOVE PLUS: LAMINAR/TURBULENT TRANSITION TURBULENCE DISSIPATION	10^{12} TO 10^{15}	3 MILLION TO 3 BILLION CLASS VI

Speed and memory requirements for computational aerodynamics are compared with several existing and planned computers in figure 4. Computers large enough to provide solutions to the Reynolds-averaged Navier-Stokes equations for the flow about a complete aircraft are expected to be available before the end of this decade. That should mark the time when computers will not be just a supplement to the aircraft design process, but they will be an absolute necessity for a country to remain competitive in meeting economic and performance requirements.

An expert group of computational fluid dynamicists, computer and wind-tunnel technologists, and airframe and engine designers were brought together by the National Research Council to assess the impact of computations on the traditional role of ground test facilities over the next 15 years (1983-1998). They concluded that, over the 15-year period considered, the capabilities of the computer as a tool for aircraft and engine design will increase substantially, the unit cost of computation will drop by three orders of magnitude, and the type of testing will change, but there will be no marked change in the requirement for using ground test facilities.[1] This latter conclusion stems from the finding that computation and experimentation play somewhat different roles and have different strengths and weaknesses. On the other hand, computations give considerably greater detail of a flow field than is possible in any wind tunnel. They provide a capability for configuration optimization and for determining the effect of configuration changes before commitment to model construction and testing is made. In addition, computations can provide an alternate source of information often needed to interpret experimental results or to extend them to conditions not obtainable in ground test facilities. On the other hand, ground test facilities provide a ready source of integrated flow information since forces and moments are directly obtained by wind-tunnel balance measurement. In summary, computations and experiments will be used in a complementary rather than competitive mode, and the end result will be products which will maintain the country's preeminence in civil and military aeronautical fields.

Aerothermodynamics

Aerothermodynamics is the extension of aerodynamics into very-high-speed flight regimes where there are significant thermal effects between gas and solid surfaces. As the motion of a vehicle increases to hypersonic speeds (Mach number greater than about 5), shock-heated molecules in the flow exhibit excitation into higher internal degrees of freedom (vibration, rotation, and electronic) and chemical reactions begin to occur. This gives rise to the onset of convective heating and eventually to dissociation of the gas molecules surrounding the vehicle. Additional gas species such as NO, N, and O are formed if the vehicle is flying in air. At even higher speeds, ionization begins causing the appearance of N^+, O^+, and N_2^+. In many instances, the gases will become sufficiently hot to radiate at intensity

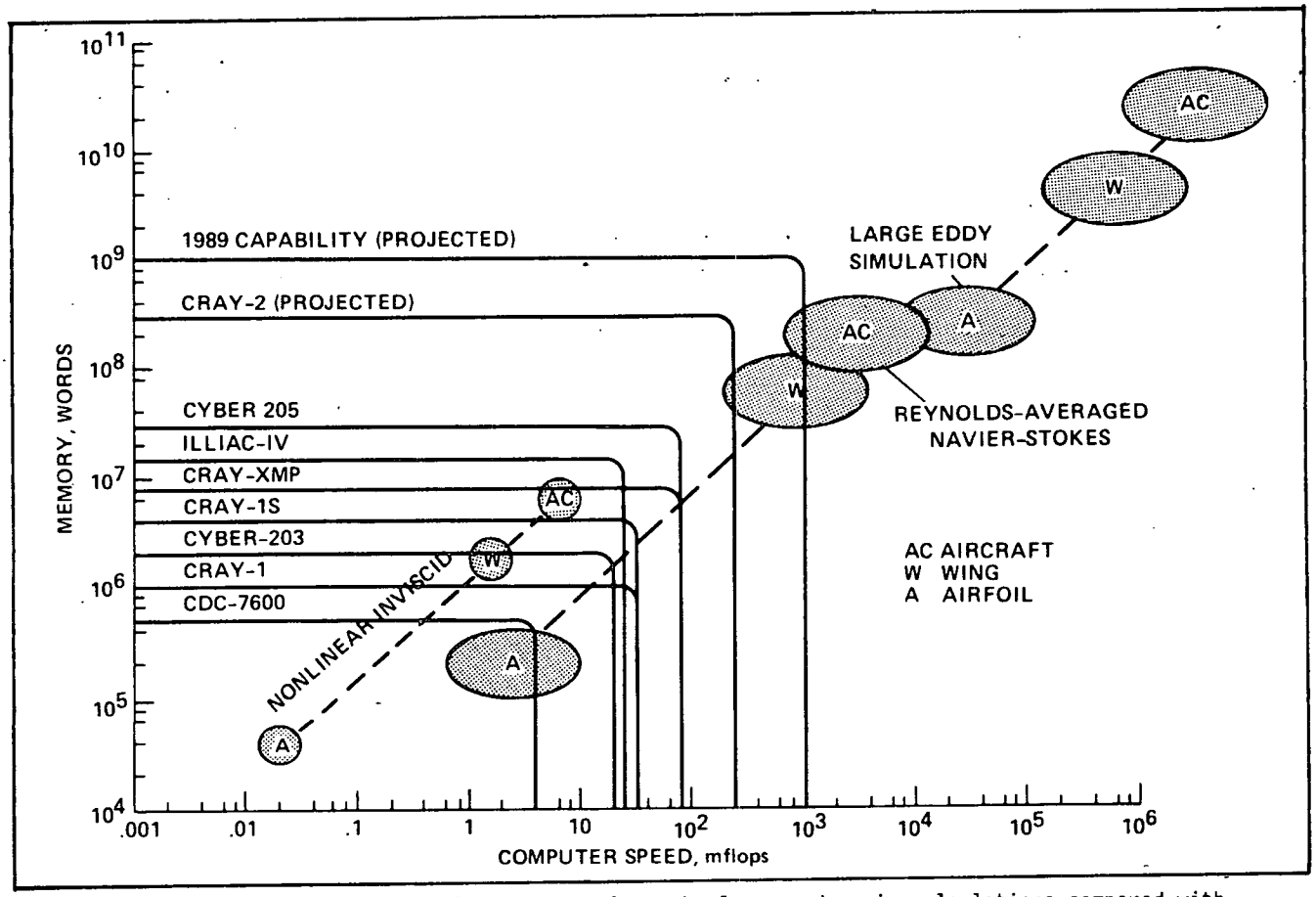


Figure 4.- Computer speed and memory requirements for aerodynamic calculations compared with the capabilities of various machines; 15-min runs with 1985 algorithms.

levels giving rise to the requirement to provide thermal protection against both radiative and convective heating to assure vehicle survival. The gas is generally in thermochemical equilibrium for flight at altitudes below about 40 km in the Earth's atmosphere, but at higher altitudes, the physics are further complicated by having to treat the flow chemistry with finite rates of reactions. All of these real-gas effects must be taken into account to accurately obtain pressures, forces, and heat loads experienced by a vehicle flying at these hypervelocities. Computational treatment of flows involving aerothermodynamic phenomena places greater demands on computer speed and memory than for the treatment of aerodynamics under perfect-gas assumptions because of the need for a more detailed description of the equation of state. These demands are offset, to some extent, by the fact that vehicles designed to fly at hypervelocities generally have simpler geometries than conventional aircraft so that fewer grid points are required to discretize the computational domain.

The impact that modern computers will have on the field of aerothermodynamics can be illustrated by considering the approach contemplated for the design of the next-generation Space Shuttle and comparing it to that used for the existing Shuttle, which was developed in the 1970s. During that decade, the computational approach was in its infancy and computers were not powerful enough to treat the complete vehicle, even with codes based on perfect-gas assumptions. The Shuttle was designed by testing many competing configurations in ground test facilities such as wind tunnels and arc-heated plasma jets. This testing program extended over a period of more than 10 years and it required over 50,000 hours of time in wind tunnels alone. Clearly, this was a highly successful program, but the approach was expensive in terms of time and added weight required to protect against uncertainties caused by not being able to precisely simulate all of the physical phenomena in the ground test facilities. It is likely that the design of the next-generation Shuttle will begin with the calculation of the performance characteristics of many possible configurations using codes based on ideal-gas assumptions. This will guide the selection of vehicle shapes that satify low-speed performance requirements while showing the promise of meeting high-speed requirements. At this point parallel experiments and

computations will begin to investigate likely areas of concern related to real-gas effects. Once the real-gas codes are validated by comparison with measurements made at conditions that can be obtained in ground test facilities, they can be used with some degree of confidence to further narrow the number of candidate configurations prior to exhaustive, time-consuming testing. Thus, the burden of "cut and try" optimization will be assigned to the computer and the results will be checked by measurement for those portions of the flight envelope accessible by ground test facilities.

Computer speed and memory requirements for treating real-gas flows about simple shapes typical of planetary probes, ballistic missiles, and orbital transfer vehicles, which will use aerobraking in the Earth's upper atmosphere to replace rocket-motor-assisted maneuvering, are shown in figure 5. Estimates are shown for increasing levels of physical complexity ranging from laminar-flow, ideal-gas considerations to the treatment of chemical and thermochemical nonequilibrium effects with radiation and turbulence physics included. The results are based upon using 1985 algorithms, the Reynolds-averaged form of the Navier-Stokes equations, and 15-min solution times.

Aeroassisted orbital transfer vehicles (AOTVs) will fly in a regime in which all of the physical complexities except turbulence are expected to be prominent. The results in figure 5 imply that about 40 hours of CPU time will be required to calculate the flow about this class of vehicle using machines having a speed of 1000 MFLOPs and a memory of 256 million words. This is not considered to be an excessive amount of computing time to provide information vital to the design of an optimized vehicle, but not readily accessible through ground test.

Chemistry

Properties of matter can be calculated by solving the Schrödinger equation, which is the fundamental equation of quantum theory. Interaction energies between species and most physical properties of interest can be calculated for systems composed of as many as 100 atoms using current algorithms and computers. With these

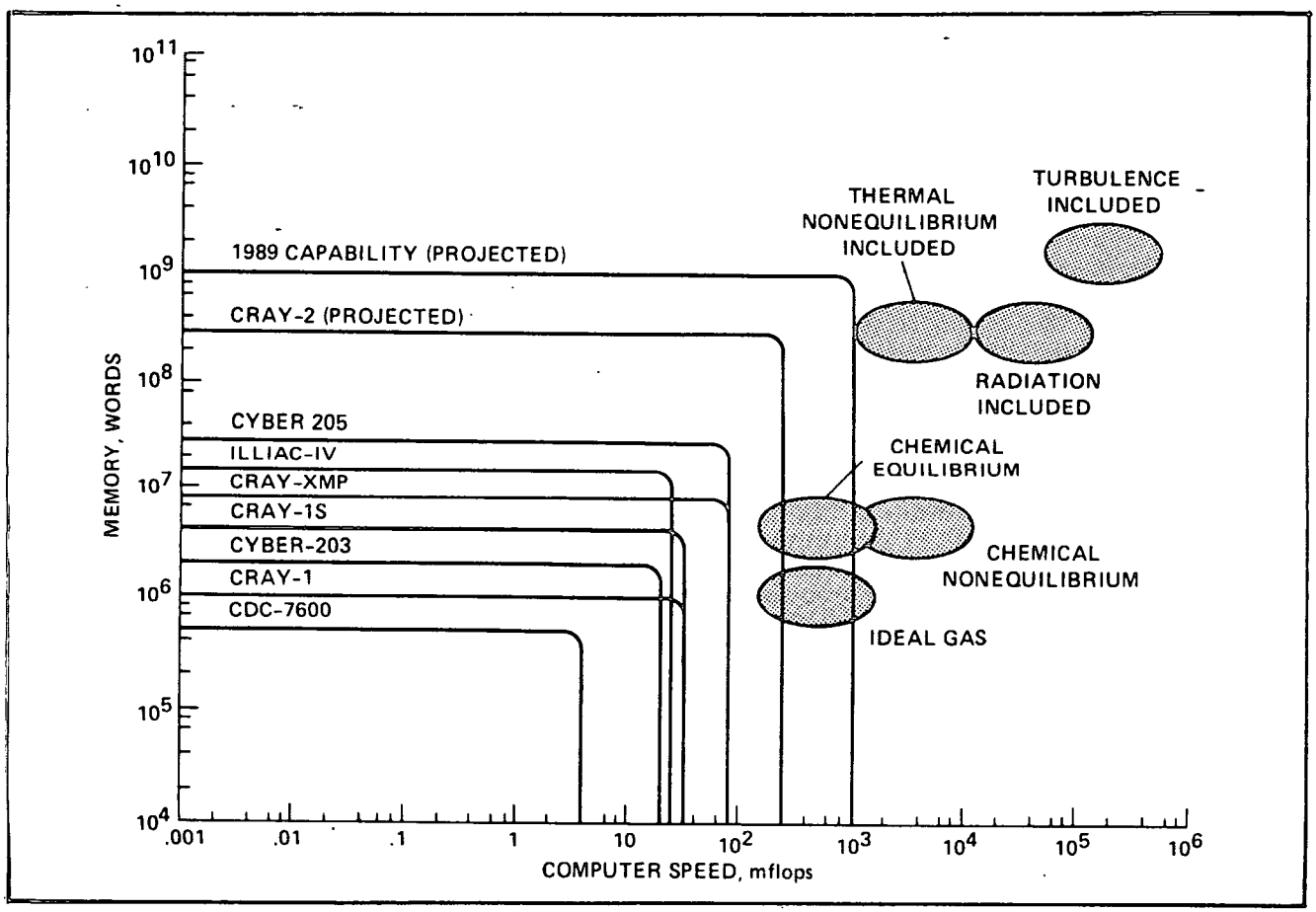


Figure 5.- Computer speed and memory requirements for aerothermodynamic calculations pertaining to planetary probes, ballistic missiles, and orbital transfer vehicles compared with the capabilities of various machines; 15-min runs with 1985 algorithms.

interaction energies, classical and semiclassical mechanics are used to compute collisional properties, including rates of chemical reactions. This field of research is known as computational chemistry, and its applications are as broad as those of chemistry itself. The discussion herein will be limited to problems dealing with atmosphere-entry physics and aerospace materials.

The solution of Schrödinger's equation is usually obtained by a basis-set expansion. Molecular orbitals for the system of interest are written in terms of a set of atomic basis functions, and the expansion coefficients are obtained by minimization of the total electronic energy for the molecule or atomic cluster; the larger the basis set, the more accurate the results. When the energy minimization is accomplished with the total number of electrons in the system assigned in only one way to the molecular orbitals, the result is known as a self-consistent field (SCF) solution. For some systems and some properties, these results are sufficient, as will be discussed later. SCF solutions usually give information on only the lowest or ground electronic state of a system, but this information often is all that is of interest. The SCF solution does not account for the fact that electrons tend to avoid each other in their motions about the nuclei of the molecule. To account for this, the electrons are allowed to occupy the molecular orbitals in many different ways following quantum mechanical rules. This approach is called configuration interaction (CI), and the resulting electronic energies and wavefunctions from whence many properties of isolated systems are derived often are as accurate as those from high-quality laboratory experiments. In general, properties of excited electronic states require the use of the CI approach and this can be computationally expensive.

Atomistic simulation of material properties utilizes the interaction energies obtained either from the quantum chemical solutions or those deduced from measurements. For the former case, quantum calculations are first conducted on an atomic cluster such as 50 or more metal atoms with and without the presence of gaseous impurity species. Effective interatomic forces accounting for two- and three-body interactions are deduced from the potential energy surfaces for the small cluster. These forces are then used as inputs for classical-dynamical-theory calculations for gas-material interactions by considering an ensemble of 10,000 atoms at a point of interest in the material with external (macroscopic) forces connected to the discrete atoms via a set of finite elements.

Computational chemistry currently is being used to provide information required for the calculation of the aerothermodynamic behavior of the AOTVs. For example, the electronic transition probabilities for N_2^+, which contribute importantly to the radiative emission from the hot air, and the rates for the $N^+ + N \rightarrow N + N^+$ charge-transfer reaction, which is an important convective heat-transfer mechanism for nonequilibrium airflows, are being provided by computation. Results based on CI studies are being obtained with reliability levels comparable to high-quality measurements. In addition, emission spectra for the various other radiating species in the flow are being calculated and provided as input to the flow-field codes to predict emission, absorption, and radiative heating. Finally, chemical reaction rates are being computed from first principles by using interaction energies or potential energy surfaces resulting from the solution of the Schrödinger equation at all possible values of the interatomic coordinates. This process involves the simulation of reactive trajectories by solving Hamilton's equations of motion. Many solutions, corresponding to different Boltzman distributions over initial vibrational states (vibrational temperatures) and approach conditions (translational temperatures), are simulated, giving rise to computed reaction cross sections. By averaging these cross sections over the appropriate vibrational distribution, effective rates of reactions can be determined. For endothermic reactions, this can lead to more than an order of magnitude difference in the rate constant, depending upon the degree to which the vibrational temperature is out of equilibrium with the translational temperature (e.g., translational temperature of 10,000 K and vibrational temperature of 4000 K).

Another application of computational chemistry is related to the development of new advanced polymers. Many physical properties of polymers depend upon their segmental motions, such as rotations of CH_3 groups about chemical bonds. For example, glass transition and toughness of structural polymers can be understood in terms of such changes on the molecular level. Experiments have revealed a wealth of information on these motions, but the interpretation of the data is often very difficult. An example of how computations can assist with this interpretation involves a recently conducted study of polymethylmethacrylate PMMA, the clear material commonly used for aircraft windows. Torsional potentials for appropriate rotations were computed using a modest basis-set expansion and SCF wavefunctions. The calculated torsional barriers to rotation fell in three ranges and agreed in absolute value with previous measurements. However, assignment of the barrier heights to internal motions deduced by the experimentalists was found to be incorrect. Guidance provided by the calculations permitted appropriate corrections to be made for interpretation of the measurements. This example illustrates the value of a combined experimental and computational

approach in the program to develop advanced polymeric materials.

The computational work is having a profound effect on the approach being taken by surface-physics experimentalists. Once it became clear that the computations would be capable of providing detailed and reliable information on the physiochemical properties of small atomic clusters, experimentalists set the goal of measuring these data for very small clusters supported on "inert" substrates. This has recently been accomplished at Ames, and measurements have been made on clusters with as few as six atoms. In other laboratories, clusters of two, three, and more heavy atoms have been seen in free jets and trapped in rare-gas matrices. Thus, the people who are conducting the experiments and those who are using computers are now working on the same "turf" in the field of small particles, and many beneficial results in interpretation have already passed between them. It is expected that these exchanges will increase with time and result in many more unforeseen benefits.

Examples of computer speed and memory requirements for three areas of computational chemistry are shown in figure 6. The first area relates to the gas-phase molecular properties required for nonthermochemical equilibrium aerothermodynamic studies of AOTVs. Radiative properties of molecules such as diatomic nitrogen, for a single molecular geometry, can be calculated on existing computers in less than 15 minutes, although about 50 such calculations are required to provide all of the necessary information. Intermolecular potential surfaces for two diatomic nitrogen molecules, in both ground or electronically excited states, require either substantially larger computers or longer runs since each point on a potential surface requires more than 15 minutes of computing time on today's computers and several hundred points are required to define a surface. Nevertheless, the cost of this amount of computer time is still less than that for measuring required rates of chemical reaction for the four-atom complex in the ground and excited states. Requirements for polymer research are illustrated by the data for vibrational and torsional potentials for polymethylmethacrylate PMMA in the monomer, dimer, trimer, and three-cluster trimer form. Again, these calculations are very lengthy on today's computers, but information is

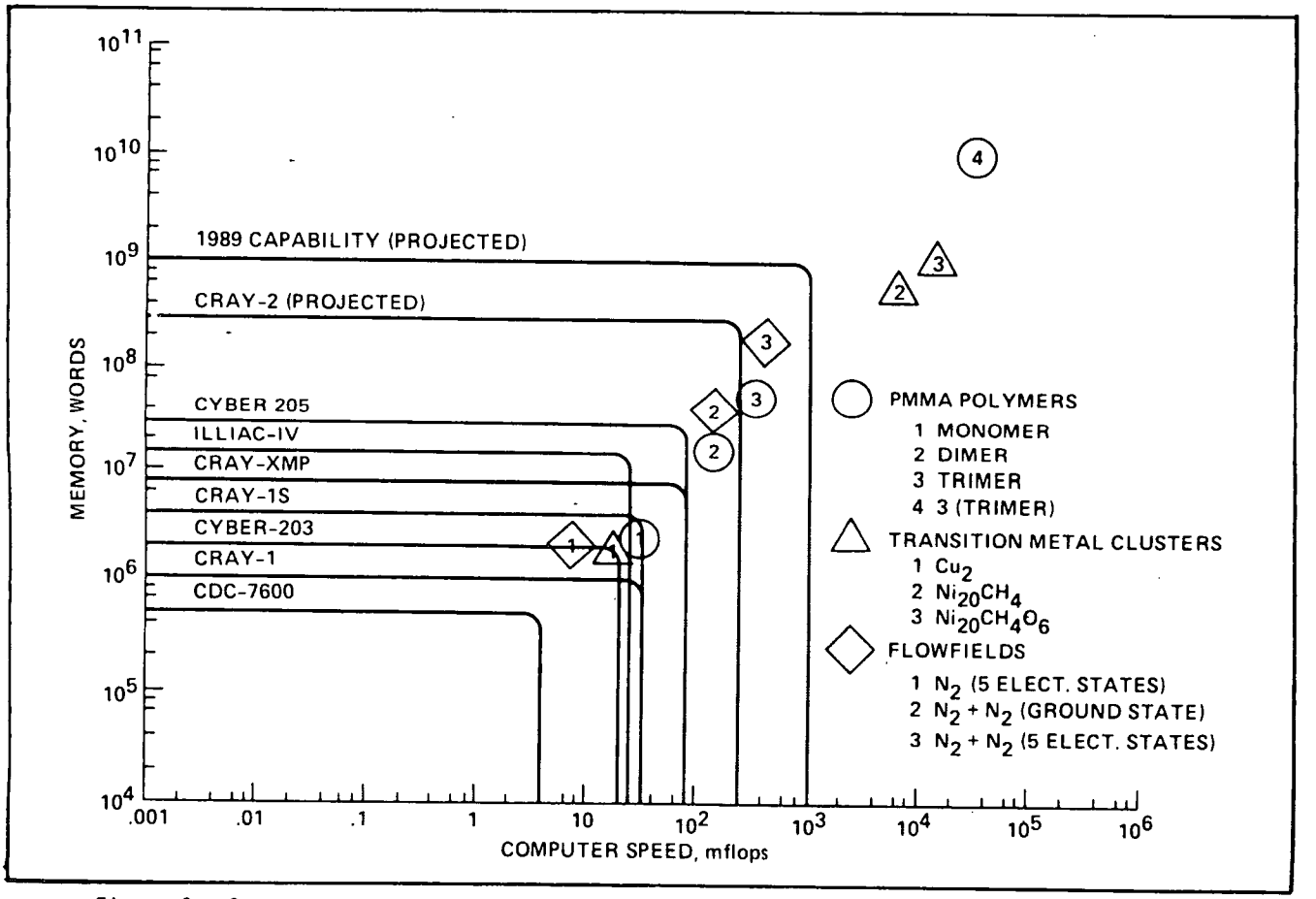


Figure 6.- Computer speed and memory requirements for three areas of computational chemistry compared with the capabilities of various machines; 15-min runs with 1985 algorithms.

provided that cannot be obtained from experiment. Finally, some estimates of computer speed and memory requirements are shown for work on small clusters of metal atoms with focus on developing an understanding of how metals interact with gases. The data for Ni_{20}, with an without gaseous molecules present, correspond to one of several hundred calculations needed to predict the physiochemical properties of small clusters of atoms supported on substrates. Information of this type is pertinent to the understanding of catalysis. It is clear from the results presented in figure 6 that computers much more powerful than those available now will be required to satisfy the future needs of the computational chemist, although considerable information of great value can be obtained economically with existing machines.

As an example of the cost effectiveness of computations in this field, consider the determination of an important electronic transition moment for the C_2 molecule. A summary of 12 different laboratory measurements made prior to 1975 showed that values reported for this transition moment differed by a factor of about 6.[4] The cost of obtaining these measured data is conservatively estimated to be about $600,000. This estimate assumes that each of the 12 experiments cost $50,000 to perform. Today, one computational chemist, spending 3 months and using about 5 hours of time on a Cray-XMP computer ($10,000), can obtain all of the transition moments between the eight lowest states of the C_2 molecule, to within ±15%, and values for the bond-dissociation energy accurate to within 0.1 eV. Fifteen years ago there was no alternative to using shock tubes or similar experimental devices to obtain data of this type; now, the computer is a cost-effective substitute.

Atmospheric Sciences

Scientific studies of planetary atmospheres are becoming heavily dependent on the use of large-scale computers for performing complex data analyses, archiving large data sets, and acting to model the physical behavior of the atmospheres over time. Models of the atmosphere provide a framework to organize knowledge, define and interpret measurements, predict both short- and long-term weather and climate, and predict environmental impacts resulting from numerous natural and human-induced perturbations. Equations governing the physical phenomena in planetary atmospheres are similar to those describing the fluid dynamics of flows about aircraft, but they must also include the effects of multiphase systems (gas/liquid), radiative transfer, and in some cases, chemical reactions.

Computers are not yet large enough to resolve all of the time-dependent physical processes occurring globally within a planetary atmosphere without the use of approximations and empirically derived information. Therefore, there is a strong interplay between computation and measurement. For example, models of stratospheric aerosol physics and chemistry have been used to guide the selection of instruments and to define critical measurements. Once measurements are made, the models are tested against these data to identify areas of data and model deficiencies. After deficiencies are corrected, the models can be used to predict the consequences of phenomena which are not directly observed, such as those resulting from clouds from volcanic eruptions or meteor impacts in the distant past. Another example involves the use of radiative transfer codes to interpret and extend measurements of the light fields in the atmosphere. Measured light values can be reduced to cloud temperatures and radiative heating rates can be related to atmosphere absorber concentrations.

Examples of computer speed and memory requirements for several selected areas of computational atmospheric sciences research are shown in figure 7. Current supercomputers are adequate for treating three-dimensional localized climate and limited-domain atmospheric models. Much larger machines still are required to extend climate modeling to include complex chemistry, increased range, and greater resolution, and to perform short-term, high-resolution, three-dimensional atmospheric simulations with exchange processes occurring between the stratosphere and troposphere. However, the threshold of required computer power clearly has been crossed for disciplines in the atmospheric sciences.

Astronomy and Astrophysics

Many important insights in modern astronomy have been obtained through large-scale computation. Astronomical phenomena typically combine complex interplays of several physical processes with strong nonlinear effects. Hostile or unattainable environments preclude laboratory studies, and many processes take millions of years to complete. Computation provides a major hope for sorting out and understanding such interacting processes.

The complexities of astronomical phenomena combine with greatly improved observational data to broaden the scope of problems that demand attention and to sharpen the detail sought in interpreting observations. Along with significantly improved accuracy and greatly increased data rates on more traditional observations, qualitatively new kinds of data from space-based observatories and imaging detectors yield a flood of data and introduce new kinds of problems that can be interpreted or solved only with the aid of computers.

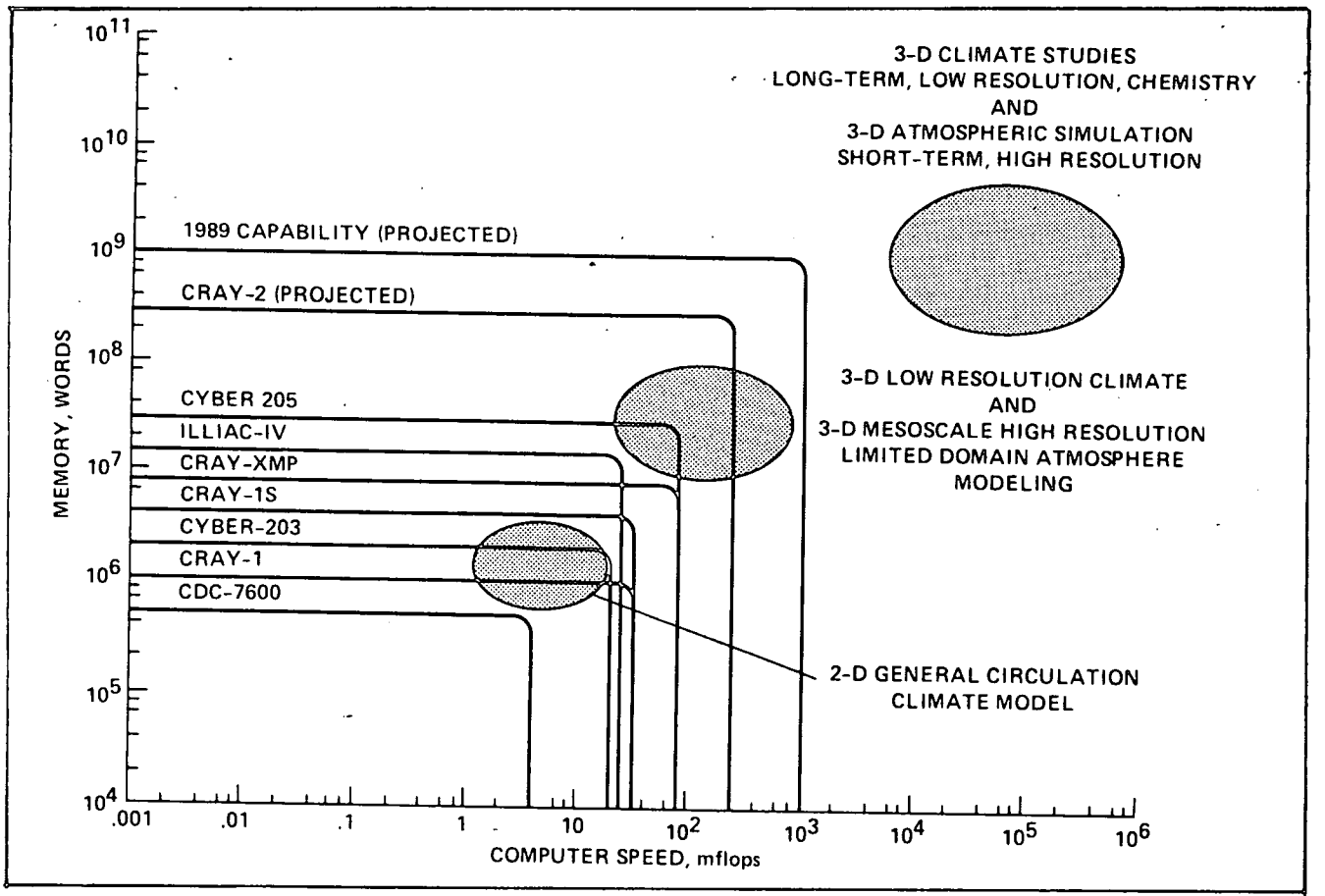


Figure 7.- Computer speed and memory requirements for selected areas of atmospheric sciences research compared with the capabilities of various machines; 15-min runs with 1985 algorithms.

Examples of the use of computers in astrophysics, involving galactic dynamics, are illustrative of current research. Dynamics of galaxies, in which galaxies with three-dimensional shapes that had traditionally been assumed were found to be dynamically unstable when treated by means of numerical experiments conducted with computers. These results led to the suggestion that spiral galaxies are embedded in massive dark "haloes" that are themselves stable and within which the galaxy would be stable. This suggestion stimulated numerous observational studies and now seems to be confirmed by observations of rotation fields in spiral galaxies. Dynamical flow fields in barred spiral galaxies and the shapes and dynamics of elliptical galaxies are other basic properties whose character has been learned from numerical experiments.

These examples illustrate how the computer can be used as an exploratory tool to uncover the dominant physics that govern observed astrophysical phenomena or, even more basically, to study qualitative properties or global stability. The thread of continuity that underlies these examples is the fact that reality is complicated and it cannot be simulated without including all of the relevant physics. Computers now permit increasingly detailed modeling which can result in wholly new understandings. This is a pattern that, through stellar studies, has been repeated over and over again in investigations ranging from the red giants to stellar pulsations, interstellar chemistry, protostellar collapse, and galactic dynamics. Other problems that require the largest and fastest available computers to treat are black hole dynamics, star formation, galactic chemical evolution, magnetic fields and plasmas, radio sources and jets, and supernovae.

Examples of computer speed and memory requirements for computational astronomy and astrophysics research are shown in figure 8. Machines of the CDC 7600 class made possible the solution of two-dimensional models of galactic dynamics and star formation. Current machines permit the move to three dimensions. Even larger machines will allow the simulation of galaxy formation with coupled hydrodynamics and stellar dynamics as well as star formation, including magnetic-field accretion disks and planet formation. It is clear that computers are becoming

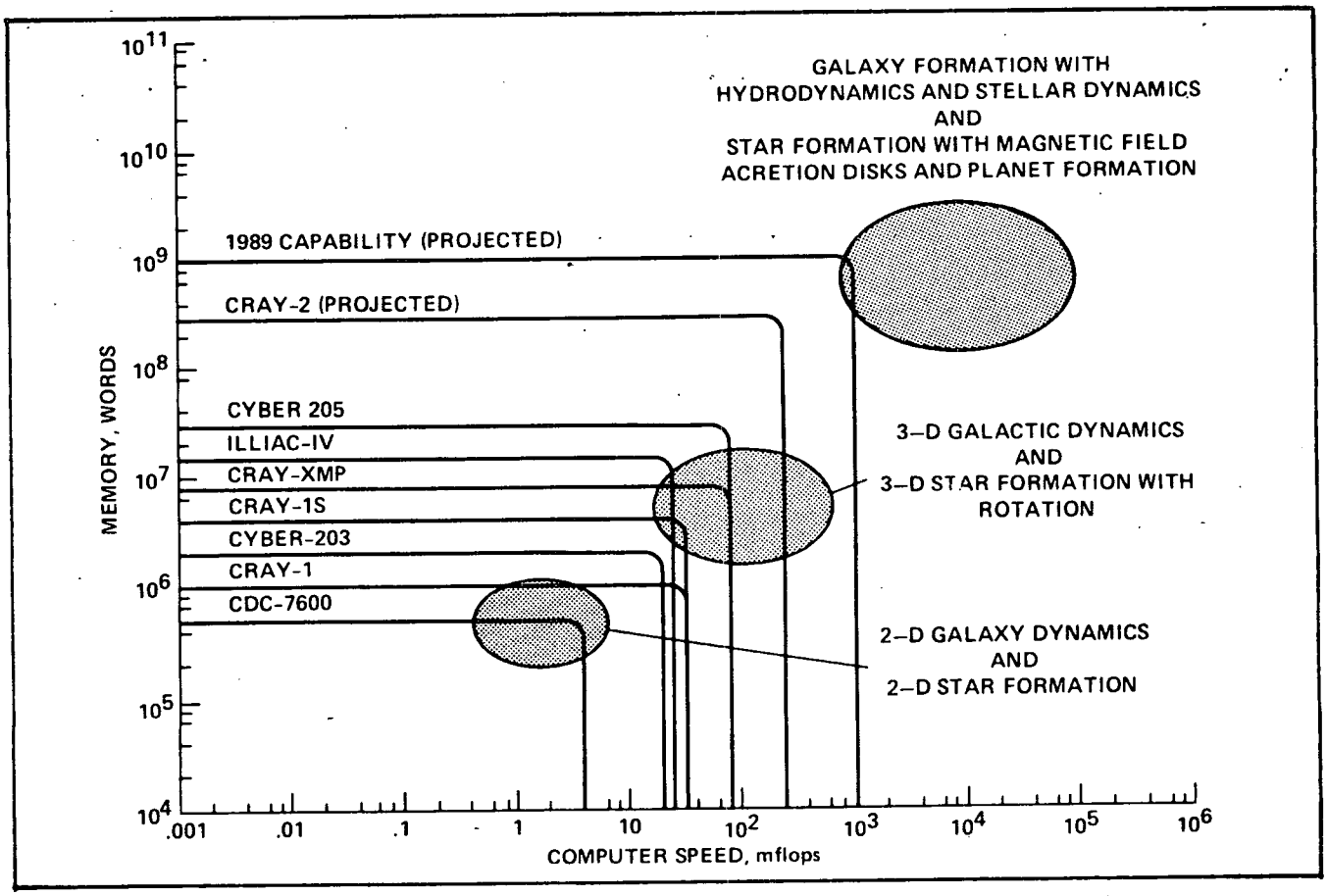


Figure 8.- Computer speed and memory requirements for computational astronomy and astrophysics research compared with the capabilities of various machines; 15-min runs with 1985 algorithms.

indispensable to research in astronomy and astrophysics, and that the need for physical experimentation to provide empirical models and validation will continue into the indefinite future.

Relative Roles of Computation and Experiment

It is clear from the examples presented that while computers are a relatively new tool for the scientist and engineer, they are already becoming indispensable to the research and development process. Each of the technical disciplines has a different threshold in terms of the computer speed and memory required for computation to be an important factor, but these thresholds are now being crossed for a wide variety of research and product-development situations. Most of the disciplines still require substantially more computer power before they are fully satisfied; however, the prospects for obtaining this power at reasonable cost in the foreseeable future are very bright.

Experience is beginning to show how this new tool will influence experimentation. In some cases, computation is a cost-effective substitute. In other cases, computation can provide information not readily amenable to measurement. In the general case, however, computation complements experiment and computers are not likely to eliminate the need for experimentation, at least for many years to come. Working together, computation and experiment can provide a greater understanding of physical phenomena and more rapid advances than relying solely on one or the other.

The relative roles of computation and experiment are beginning to change. In earlier years, experiments were performed to establish the variables governing physical phenomena and their interrelationships, and computations were made after the fact either to validate computational methodology or to provide additional understanding of the experimental results. Now, computations are being used to establish these relationships beforehand, and experiments are being conducted either to validate the computations, to provide refinements, or to assist with the modeling of terms in the governing equations that are still too complex to solve from first principles. The development of experiments and the interpretation of results now is heavily dependent on computation. Therefore, the experimentalist of the future should receive at least some training in

computation to be able to either to perform the computations required to define a good experiment or to be in a position to communicate with others who might be responsible for the computations.

The need for persons trained in large-scale scientific computation in all of the disciplines is growing very rapidly. To satisfy this need, universities, research laboratories, and industrial organizations involved with product development need greater access to supercomputers and the peripheral equipment required to make them effective. This access cannot be provided at the expense of other laboratory equipment since there will be a continuing need for experimentation. In fact, the rapid improvements in computers and computational methodology are placing more stringent requirements on the experimentalists to measure more fundamental quantities with greater degrees of precision. This, in turn, is forcing the test facilities and their supporting instrumentation to become more dependent on computer technology as well. The move toward more dependence on computers is irreversible and ways must be found to adjust to this important revolution.

Conclusions

Computers are changing the approach to the conduct of research and development in many of the science and engineering disciplines. The rapid and continuing growth of their speed and memory enable them to be used to solve, with high degrees of precision, increasingly complete forms of first-principles governing equations. They are, in fact, becoming an attractive complement to, and in some cases a cost-effective substitute for experiment.

Differing complexities of the underlying physics in the various disciplines place different requirements on the power that a computer must have to solve the governing equations in a reasonable amount of time. Thus, each discipline has a threshold of computer power that must be exceeded before the computer becomes an effective tool in that discipline. Research and applications in the disciplines for which this threshold has been crossed, such as fluid dynamics, aerodynamics, aerothermodynamics, chemistry, atmospheric sciences, astronomy, and astrophysics, now are as dependent upon the computer as they are upon physical observation and testing. In fact, many of the experiments in these disciplines are defined by the computations.

The irreversible move toward more dependence on computers is placing new requirements on both human and financial resources. There is a growing need for people with formal training in the use of supercomputers. Universities, industrial corporations, and government laboratories must have increased access to these machines to meet this need. In addition, equipment must be upgraded frequently to remain current. The complementary rather than competitive nature of computation and experiment also means that experiments cannot be abandoned in favor of computation. Thus, laboratory equipment also must continually be supported and upgraded, at least for the foreseeable future.

Computers are an indispensable new tool for the scientist and engineer. This tool will continue to become more powerful with time as hardware, software, and people skilled in their use mature. The effective combination of computation and experiment will contribute greatly toward the maintenance of a competitive position in world affairs.

References

1. National Research Council: Influence of Computational Fluid Dynamics Upon Experimental Aerospace Facilities: A Fifteen Year Projection. National Academy Press, 1983.

2. Peterson, Victor L.: Impact of Computers on Aerodynamics Research and Development. IEEE Proc., vol. 72, Jan. 1984.

3. Chapman, Dean R.: Computational Aerodynamics Development and Outlook. AIAA Paper 79-0129, Jan. 1979.

4. Cooper, D. M., and Nichols, R. W.: Measurements of the Electronic Transition Moments of C_2-Band Systems. JQSRT, vol. 15, Jan. 1975.

VICTOR L. PETERSON

Mr. Peterson joined the staff of the NASA Ames Research Center as an Aeronautical Research Scientist in 1956, upon graduation from Oregon State University with a B.S. degree in aeronautical engineering. He subsequently received the M.S. degree in aeronautics and astronautics sciences from Stanford University in 1964, and the M.S. degree in management in 1973, upon completion of study as an Alfred P. Sloan Fellow at the Massachusetts Institute of Technology. In 1984, he was appointed to his present position as Director of Aerophysics. Principal programs within his Directorate include computational and experimental fluid dynamics, high-speed aerodynamics, computational chemistry, atmosphere entry technology, computer science, Space Station automation technologies and the Numerical Aerodynamic Simulation (NAS) Program. The Directorate operates many of the Center's major facilities including wind tunnels and the large-scale scientific computers. He has written over 40 technical papers and reports in the fields of fluid and flight mechanics and on the use of supercomputers in science and engineering.

Dr. Arnold came to Ames Research Center as a Research Scientist in 1962 with a B.S. degree in Engineering Physics from the University of Kansas. He received a M.S. in Aeronautics and Astronautics from Stanford in 1967 and a Ph.D. in Molecular Physics from York University (Toronto) in 1972. He has served as Chief of Ames' Computational Chemistry and Aerothermodynamics Branch (and predecessor, Physical Sciences Branch) since 1979. Dr. Arnold currently manages work in Computational Chemistry, Surface Physics, Aerothermodynamics and Computer Science relating to large-scale scientific computing. He has published 25 papers in these areas involving both theoretical and experimental studies.

HISTORY OF THE NUMERICAL AERODYNAMIC SIMULATION PROGRAM

Victor L. Peterson and William F. Ballhaus, Jr.
NASA Ames Research Center

ABSTRACT

NASA's Numerical Aerodynamic Simulation (NAS) program has reached a milestone with the completion of the initial operating configuration of the NAS Processing System Network. This achievement is the first major milestone in the continuing effort to provide a state-of-the-art supercomputer facility for the national aerospace community and to serve as a pathfinder for the development and use of future supercomputer systems. The underlying factors that motivated the initiation of the program are first identified and then discussed. These include the emergence and evolution of computational aerodynamics as a powerful new capability in aerodynamics research and development, the computer power required for advances in the discipline, the complementary nature of computation and wind tunnel testing, and the need for the government to play a pathfinding role in the development and use of large-scale scientific computing systems. Finally, the history of the NAS program is traced from its inception in 1975 to the present time.

INTRODUCTION

The Numerical Aerodynamic Simulation (NAS) program is an outgrowth of the discipline of computational fluid dynamics. However, the NAS system is now recognized to be an important facility for advancing all of the computationally intensive aerospace disciplines and for serving in a pathfinder role for the development and use of future supercomputer systems. In fact, the NAS Program began to influence both discipline-oriented users and developers of supercomputers even before the system was first assembled. The NAS has drawn national attention to the importance of scientific computers to the country's technology base and has served as a focal point for the large-scale scientific computing community.

The NAS program will provide a leading edge computational capability to the national aerospace community. It will stimulate improvements to the entire computational process ranging from problem formulation to publication of results. The program has been structured to focus on the development of a complete computer system that can be upgraded periodically with minimum impact on the user and on the ever increasing inventory of applications software. The NAS system, in its initial operating configuration, is already serving over 200 users nationwide at over 20 remote locations. These numbers will continue to increase as the system matures to its extended operating configuration including two powerful supercomputers, all of the necessary supporting equipment, and well established communications links.

The objectives of this paper are twofold: 1) to identify the factors that led to the initiation of the NAS Program, and 2) to review the evolution of the NAS Program from its inception in 1975 to the present time. Included in the discussion are brief reviews of the evolution of computational aerodynamics, computer requirements for future advances, the complementary roles of computation and experiment, and the historical role of the government in the development and use of large-scale scientific computing systems.

FACTORS MOTIVATING THE NAS PROGRAM

The underlying motivations for the NAS program are a composite of four principal factors: 1) the emergence and evolution of computational aerodynamics as a powerful new capability in aerodynamics research and development; 2) the demands that this relatively new discipline places on computer systems; 3) the use of computation as a complement to wind-tunnel testing; and 4) the long standing, recognized need for the government to play a pathfinding role in the development and use of large-scale scientific computing systems. Each of these factors will be briefly discussed prior to describing the evolution of the program.

Emergence and Evolution of Computational Aerodynamics

Electronic computers were used to assist with aerodynamic analyses ever since they became available to the aeronautical researchers in the 1950s. Prior to 1970, aerodynamic analyses were limited primarily to the solution of the linearized inviscid flow equations and to the equations governing the behavior of the viscous boundary layer adjacent to an aerodynamic surface. Computers of the IBM-360 and CDC-6600 class permitted these equations to be solved for the flows about idealized complete aircraft configurations, but only for situations where the flows were everywhere either subsonic or moderately supersonic and everywhere attached to the surfaces over which they passed. Some attempts were made to include the nonlinear terms in the inviscid flow equations and solve for transonic flows about airfoils, but

these were limited to the very restrictive situations of either nonlifting airfoils or airfoils with detached bow shock waves.

The year 1970 marked the beginning of a series of advances in computational aerodynamics that would not have been possible without computers. The first major advance in solving for the nonlinear transonic flows about practical lifting airfoils with embedded shock waves was reported in the literature by Magnus and Yoshihary (1970). Subsequent milestones in the development of the technology for treating the nonlinear inviscid equations, and enabled only by the computer, are shown in figure 1. By about 1973, solutions for wing-body combinations treated with the steady-flow, small-disturbance equations were being published. Results of the first treatment of unsteady flows about airfoils appeared in the literature by Ballhaus, Jr., et al. (1975), and the first flutter analysis for a swept wing was published about 6 yr ago by Borland and Rizzetta (1981). Research on the aeroelastic behavior of wings is still limited by the performance of currently available computers to the treatment of the equations governing inviscid flows. These equations, with corrections for boundary-layer effects, are still used extensively for a wide range of aerodynamic problems. However, the really important problems facing the designers today require the use of the Reynolds-averaged, Navier-Stokes equations, both with and without the inclusion of the additional equations governing real-gas chemistry.

Milestones in the use of the Reynolds-averaged, Navier-Stokes equations for treating compressible viscous flows are shown in figure 2. These equations account for most of the physics of interest in fluid-dynamic flows. The process of time-averaging the Navier-Stokes equations over a time interval that is long relative to turbulent eddy fluctuations, yet small relative to macroscopic flow changes, introduces new terms representing the time-averaged transport of momentum and energy, which must be modeled using empirical information. Very powerful computers are required for simulations with this level of approximation, but the potential advantages over the inviscid equations are enormous. Realistic simulations of separated flows and of unsteady viscous flows, such as buffeting, will become commonplace as the ability to model the turbulence terms matures. Combined with computer-optimization methods, these simulations should make it possible to develop designs optimized for various missions while adhering to practical constraints such as available engine power and sufficient fuel volume to meet range requirements. Landmark advances include the investigation of a shock-wave interaction with a laminar boundary layer reported by MacCormack (1971), the treatment of high-Reynolds-number transonic airfoil flows by Deiwert (1974), the first turbulent flow over a lifting wing by Mansour (1984), and the first turbulent flow over a realistic fighter configuration at angle of attack by Flores et al. (1987). Relatively large amounts of computer time are still required for the application of these equations to practical problems, but advances in technology continue to improve computational efficiency.

Figure 3 displays a perspective on the effect that increasing computer power has had on computational aerodynamics in a practical engineering sense. Presently available machines are adequate for calculating the flows about relatively complex configurations with the inviscid-flow equations. However, the type of information derived from the computations is limited (e.g., no total drag and no effects of flow separation). The viscous-flow equations, being more complex and requiring finer computational meshes, demand substantially greater computational power to solve. Thus, the types of problems that can be solved with a given computer are necessarily less complex. In effect, a designer has to make the choice between treating simple configurations with complex physics or treating complex configurations with simple physics. Yet, in both inviscid- and viscous-flow situations, each new generation of computers has resulted in advances in the value of computational aerodynamics as a design tool. The discipline will begin to mature when both complex configurations and complex physics can be treated simultaneously with a reasonable amount of computer time.

Computer Requirements

Computer requirements for computational aerodynamics can be related to the four major levels of approximation to the Navier-Stokes equations that were identified in the work by Chapman (1979). Each level of approximation resolves the underlying physics to a different degree, provides a different level of understanding, and requires a different level of computer capability. Table 1 and the works of Chapman (1979) and Peterson (1984) discuss in some depth these approximations, their capabilities to solve problems associated with aircraft aerodynamics, and the computer requirements to solve them in a reasonable amount of time (about 15 min) for flows about relatively complete aircraft configurations. Computer requirements are expressed in terms of the power of a Class VI machine, which is defined here to have a processing speed of 30 million floating-point operations per second (MFLOPS) and a memory of about 8 million words. Machines of this class are widely available at the present time. Computer requirements increase with each higher level of approximation, both because more flow variables are involved and because either more panels or more grid points are required to resolve the flows to a level of detail that is commensurate with the physics embodied in the approximation. Experience indicates that the Reynolds-averaged form of the Navier-Stokes equations probably will be adequate for most design-oriented problems. The effects of all scales of turbulence are modeled in this level

of approximation; the development of appropriate turbulence models is the subject of current research by both computational and experimental fluid dynamicists. In fact, the experimentalists are being guided, to a large extent, by computational research programs which are based either on the large-eddy simulation approximation or on the use of the full Navier-Stokes equations for simple flow geometries.

Speed and memory requirements for computing the aerodyamic behavior of shapes of varying complexities are compared with several existing and planned computers in figure 4 . Computers large enough to provide solutions in 15 min or less to the Reynolds-averaged, Navier-Stokes equations for the flow about a complete aircraft are expected to be available before the end of this decade. This advance should mark the time when computers will not be just a supplement to the aircraft design process, but will be an absolute necessity to be competitive in meeting economic and performance requirements. Computers having even more power will be required in the future, however, to treat routine problems involving real-gas chemistry, the coupling of the disciplines of aerodynamics, structures, propulsion and controls, and the optimization of a complete aircraft design.

Complementary Nature of Computation and Experiment

In the early 1970s, computations were recognized by a few visionaries to have the potential for becoming an effective complement to fluid- and aero-dynamic experiments for a number of reasons. First, the physics of fluid flows could be represented by mathematical equations, and computers, beginning with the IBM 360 and the CDC 6600 machines, were becoming sufficiently powerful to solve meaningful approximating sets of these equations in a practical amount of time and at reasonable cost.

Second, wind tunnel costs and computational costs were recognized to be changing in importantly different ways. Increased complexity and broadened performance envelopes of aircraft caused the number of wind tunnel hours expended in the development of new aircraft to increase exponentially with time. In fact, this increase amounts to as much as a factor of about 1,000 over an 80 yr period (50 hr for the Wright Flyer compared to 50,000 hr for the Space Shuttle). Concurrently, the cost per hour of testing also increased by a factor of about 1,000 over the same period. Thus, wind tunnel testing costs escalated by nearly a million fold in 80 yr, while the cost of numerically simulating a given flow is shown by the data in figure 5 to have decreased by a factor of 100,000 in just 15 yr during the period from 1969 to 1984. This decrease was due to improvements in both computers and algorithms.

Third, on the one hand, all wind tunnels are known to have all or some of the fundamental limitations such as model size (Reynolds number), temperature, wall interference, model support interference, unrealistic aeroelastic model distortions under load, stream nonuniformity, unrealistic turbulence levels, and test gas (of concern for the design of vehicles for flight in the atmospheres of other planets). On the other hand, if it is accepted that the physics of fluid flows can be described precisely by mathematical equations, then the only fundamental limitations of the computational approach are the limits of computer speed and memory, and speed and memory appear to be expandable with time by many more orders of magnitude.

Finally, wind tunnels and computers each bring different strengths to the research and development process. The wind tunnel is superior in providing detailed performance data once a final configuration is selected, especially for cases involving complex geometry and complex aerodynamic phenomena. Computers are especially useful for other applications including: 1) making detailed fluid physics studies, such as simulations designed to shed light on the basic structure of turbulent flows; 2) developing new design concepts, such as swept forward wings or jet flaps for lift augmentation; 3) sorting through many candidate configurations and eliminating all but the most promising before wind tunnel testing; 4) assisting the aerodynamicist in instrumenting test models to improve resolution of the physical phenomena of interest; and 5) correcting wind tunnel data for scaling and interference errors. The combined use of computers and wind tunnels captures the strengths of each tool.

Pathfinding Role of the Government

A concern in the mid-1970s was that computer power was only marginally adequate for calculating the aerodynamics of simple aircraft shapes at cruise conditions. More power was needed to provide both for increased resolution of geometry and for including more complete flow physics in the analyses to predict performance during maneuvers and near performance boundaries. In fact, treatment of these more complex problems in an effective manner required advances not only in computing engines, but also in operating systems, languages, compilers, central storage capabilities, networking, remote communications, graphics, and user workstations. There seemed to be no assurance that the advances required to meet government needs would be provided without government stimulus. In fact, this view was reinforced by the information summarized in table 2 which shows the historical role of the government in stimulating the development of advanced computers. Every major new digital computer from the IBM 701 to the current Cray and Control Data Corporation (CDC) machines has evolved from technology developments accelerated by a government-sponsored pioneering

computer development undertaken to satisfy a driving need. The need for a superior design capability for aerospace vehicles was, and still is, a strong driver for the NAS Program.

NASA first became involved with the pathfinding role in large-scale scientific computers in a formal way when, in 1972, it joined with the Advanced Research and Development Projects Agency (now DARPA) to test the feasibility of the ILLIAC-IV computer. The ILLIAC Project was originally undertaken for the purposes of exploring the feasibility of parallel processing and advanced-computer-logic circuit technology, and researching new ideas for high-speed computer memory. When ARPA started the ILLIAC Project, their driving need was for an anti-ICBM control system. NASA's motivation for later joining in the development was, of course, the need for more computer power for the development of computational aerodynamics.

The CDC was experimenting with the STAR-100 computer at the same time the ILLIAC-IV was being tested. Only four of these machines, featuring new ideas in pipeline architecture, were produced. Three of these were obtained by Government laboratories and one was retained by CDC. Cray Research, Inc. had yet to produce a machine and IBM elected not to compete in the large-scale scientific computer market. Two other companies, Burroughs and Texas Instruments, were on the verge of discontinuing their supercomputer efforts. Technology surveys showed that computers having many times the power of the ILLIAC-IV and the STAR-100 could be developed, but the development would not happen without Government sponsorship since the market for supercomputers was still very small and limited primarily to government laboratories. In the mid-1970s, ARPA's interests had been largely satisfied with the ILLIAC-IV, and no government organization other than NASA appeared to be interested in first defining long-range requirements for supercomputers and then strongly urging their development.

The experience gained with the ILLIAC-IV project and the clear benefits derived from it provided further motivation for proceeding with a major thrust to develop an advanced computational system and the confidence that success could be achieved. Benefits from the ILLIAC-IV Project accrued in four major areas. First, in computer technology, the ILLIAC-IV was the first large machine to have multiple processors working in parallel, the first to employ emitter-coupled logic (ECL), and the first to have multilayered (12 layers) printed circuit boards designed with automated methods. Second, in algorithm technology, the existence of the machine forced the development of numerical methods for parallel processing. This new method also led to the revelation that some principles of parallel algorithms could be utilized to obtain faster execution of problems on conventional computers of that time period that could perform some functions simultaneously, such as the CDC 7600, than could be obtained using algorithms based on sequential computing concepts. Third, a deeper understanding evolved from the problems associated with large one-of-a-kind scientific computers. These problems included operating-system software costs, problems associated with applications software transportability to machines having different architectures, and a need to provide extensions to the common FORTRAN language to obtain maximum performance gains. In fact, the NASA Ames Research Center's investigators developed a language called "CFD" which enabled fluid dynamics codes to be run efficiently on the parallel-processing architecture. For problems that could be structured in parallel, the ILLIAC-IV was substantially more powerful than the other scientific computers of its era.

This advanced computer power enabled a number of pioneering advances in CFD, including the first simulation of viscosity-induced unsteady flow (buffett) about an airfoil, the first simulation of control-surface buzz, and detailed simulations of turbulent flows. The ILLIAC-IV experience provided the foundation and motivation for continuing to advance both CFD and supercomputer systems technology, which led to the conception of the NAS program.

EVOLUTION OF THE NAS PROGRAM

The potential value of the computational approach to aerodynamics research and development was clearly established by the mid-1970s. Also clear was the importance of pursuing every conceivable opportunity for improving aerospace vehicle design tools to maintain a leadership position in the intensifying international competition in both the commercial and military aircraft arenas. Thus, in 1975, a small group of people associated with the computational fluid dynamics effort at the Ames Research Center conceived the NAS program as a vital underpinning of the country's future in aeronautics.

The group recognized the importance to computational aerodynamics of a sustained effort to increase computer power as rapidly as technology would allow. They also recognized the need for the government to assume some responsibility for a pathfinding role to accelerate the attainment of new milestones in computer performance.

The initial proposal called for the development of a special-purpose processor called the Navier-Stokes Processing Facility. The central processor was to have a minimum effective speed of one-billion floating-point operations per second when operating on the three-dimensional, Reynolds-averaged, Navier-Stokes equations and to have performance comparable to the best general-purpose computers when used for processing the equations of other scientific disciplines. Its main memory had to accommodate a problem data base of

31-million 64-bit words. To keep development risks low, the goal of the project was to assemble existing computer component technologies into a specialized architecture rather than to develop new electronic components. Finally, the machine had to be user-oriented, easy to program, and capable of detecting systematic errors when they occurred. The proposal was endorsed in principle by NASA management in November, 1975; then in-house studies began to gather momentum and the name of the project was changed to the Computational Aerodynamic Design Facility (CADF).

Computational Aerodynamic Design Facility Project

The first formal exposure of NASA's objectives occurred in October, 1976 when proposals were requested from industry to "perform analysis and definition of candidate configurations for a computational facility in order to arrive at the best match between aerodynamic solution methods and processor system design." These analyses were to be directed toward the selection, preliminary design, and evaluation of candidate system configurations that would be best suited to the solution of given aerodynamic flow models. Design requirements that were established for this study included: 1) the capability to complete selected numerical solutions of the Navier-Stokes equations for grid sizes ranging from 5×10^5 to 1×10^6 points and wall-clock times (exclusive of input-data preparation and output-data analysis) ranging from 5 to 15 min; 2) a working memory of 40×10^6 words; 3) an archival storage of at least 10×10^9 words; and 4) 120 hr/wk of availability to the users.

Two parallel contracts were awarded in February 1977 to develop preliminary designs for the most promising configurations and to develop performance estimates, risk analyses, and preliminary implementation cost and schedule estimates for each of the designs. During these initial studies, which lasted about 12 mo, it became apparent that the overall approach to developing the facility was sound and that performance goals could be reached with new architectural concepts and proven electronic components.

A 3-day workshop on Future Computer Requirements for Computational Aerodynamics was held at the Ames Research Center in October 1977 for the purposes of further clarifying the need for a large-scale computer system for computational aerodynamic work, for confirming that the design goals were consistent with the needs of the projected users of the facility and for validating the feasibility of meeting the requirements with emerging technology. Representatives from all of the appropriate technical communities were invited, including aircraft companies, computer companies, software houses, private research institutions, universities, the Departments of Defense and Energy, and other NASA Centers. An unanticipated large attendance of over 250 peopl confirmed the existence of broad national intere and need for more powerful computers in science and engineering. The feasibility of meeting processing speed and memory requirements was further solidified, although it was clear that the goals could only be met with a multiple-processor architecture. Projected near-term advances in electronic component performance would not permit the goals to be met with a single-processor machine. The workshop also confirmed that computer industry economics at that point in time would not support the development of large specialized processors without the infusion of government capital. The market at that time was uncertain, and it was not clear that enough machines could be sold to amortize the development costs. Finally, the aircraft industry reaffirmed the need for the proposed facility for use in solving special design problems and for serving as a pathfinder for the development and use of large-scale scientific computer systems. The workshop proceedings were edited by Inouye (1978).

An assessment of the utility of the Computational Aerodynamic Design Facility for disciplines of interest to NASA, other than fluid- and aerodynamics, was also conducted in 1977. This assessment was initiated to provide assurance that the facility would not be so highly optimized for solving the fluid dynamic equations that it would not be useful for other work. It would also provide guidance as to how the design could be altered, if required, to make it useful for general science and engineering calculations without seriously impacting its capabilities for the originally intended problems. Experts involved with research on weather and climate, structures, chemistry, astrophysics, and propulsion reviewed the proposed architectures and analyzed how the various solution algorithms peculiar to those disciplines could be mapped onto the designs. Results of the assessment confirmed the expected conclusion that the CADF would provide a powerful new capability for a broad range of problems of importance to NASA.

Numerical Aerodynamic Simulation Facility Project

After it was recognized that the facility would be used primarily for computational research rather than for routine aircraft design, the name was changed during the course of the first study contracts to the Numerical Aerodynamic Simulation Facility (NASF). Even though it became apparent after the workshop that a computational resource of this magnitude would be a valuable tool for the solution of complex problems in other technical areas of interest, aerodynamics would still be the discipline used to drive the requirements. However, before the conclusion of the first round of contracted efforts, the need for further studies with greater emphasis on a computer suitable for a broader range of disciplines was recognized.

Accordingly, 12-mo follow-on feasibility study contracts were awarded in March 1978. The results of these efforts were expected to provide data of sufficient accuracy to permit formulation of a definitive plan for the development of the facility. Several events occurred during the period of these studies which resulted in some revisions to the basic performance specifications and a deeper involvement of the user community in the project activities.

The discipline of computational aerodynamics had matured significantly in the 3 yr since the project was first conceived. New numerical methods were developed and existing methods were refined. This led to the realization that if the size of the on-line or working memory was increased to 240×10^6 words, the facility could be used not only to estimate the performance of relatively complete aircraft configurations, but also to serve as an effective tool to study the physics of turbulent flows, a subject that had eluded researchers for more than 80 years. A corresponding increase in the off-line file storage from 10×10^9 to approximately 100×10^9 words was required to accommodate the larger data sets.

A User Steering Group was formed in July 1978 to provide a channel for the dissemination of information regarding project status, a forum for user-oriented issues needing discussion, and a sounding board by which the project office could obtain feedback from future user organizations. Examples of user-oriented issues of interest were: 1) selection of user languages; 2) management policy; 3) equipment required for remote access; and 4) data protection. The User Steering Group was composed of representatives of the aerospace industry, universities, and other government agencies. The group is still active, although its name was eventually changed to the User Interface Group to reflect its current role more accurately. Organizations currently represented on the User Interface Group are shown in table 3.

The feasibility studies were completed in the spring of 1979. Each study produced a refined baseline configuration, a functional design, and rough estimates of cost and schedule. Both studies concluded that about 5 yr would be required to complete the detailed design and to develop, integrate, and test the facility. While preparations were being made to continue the contracted development process, the name of the project was changed once again to the Numerical Aerodynamic Simulator (NAS) Project.

Numerical Aerodynamic Simulator Project

A detailed plan for the design-definition phase of the activity was prepared during the winter of 1979 by the NAS Project Office, which was established at Ames Research Center earlier in the year. This plan included refining the specifications for: 1) the computing engine; 2) the support processing system; and 3) the collection of other peripherals, including intelligent terminals, graphical display devices, and data communication interfaces to both local and remote users. Two 40-week, parallel, design-definition contracts were awarded in September 1980. Upon their completion in July 1981, the contractors were awarded follow-on contracts related to further design definition. These were concluded in April 1982 when the proposals for the detailed design, development, and construction were submitted by the contractors for evaluation.

After an evaluation of the proposals, the decision was made in June 1982 to discontinue the procurement. This decision was based on evaluation findings which were that the risks involved in achieving the proposed technical objectives within the critical resouree and schedule limitations were unacceptable. Following this decision, efforts began to chart a new course of action. A reassessment was made of the needs of the user community and the evolving state of the art in computer technology. Three principal conclusions resulted from this reassessment.

First, the application and essential importance of computational aerodynamics to aeronautical research and development had grown significantly since the mid-1970s. Thus, it was deemed important to establish and to maintain a leading-edge computational capability as an essential step toward maintaining the nation's leadership in aeronautics. To achieve this goal the NAS project was to be restructured as an on-going NAS program in which significant advances in high-speed computer technology would be continuously incorporated as they became available.

Second, the supercomputer environment had changed since the inception of the NAS activity in the mid-1970s. Increased interest in supercomputing, advances in computer technology stimulated in part by the NAS Program, and the increasing threat of foreign competition changed the environment to the extent that it no longer appeared necessary for the government to directly subsidize the development of the next generation of scientific computers. These factors provided an environment permitting a more systematic, evolutionary approach toward developing and maintaining an advanced NAS computational capability.

Third, the importance of coupling advancements in the state of the art of supercomputers with advanced system networks and software architectures was recognized. This capability is necessary to accommodate successive generations of supercomputers from different vendors and to provide the capabilities needed to enhance productivity of the user. This step led to a strategy that minimizes the dependence of the entire system on single vendors and to the establishment of a strong in-house technical capability to direct the initial and ongoing development efforts.

This reassessment highlighted the importance of the pathfinding role of the NAS program. It would be particularly challenging to develop a system with components ranging from supercomputers to user workstations that could be maintained at the leading edge of the state of the art, while simultaneously providing uninterrupted service to a large community of users working on important national problems.

Numerical Aerodynamic Simulation Program

A plan for the redefined program was approved in February 1983. It included: 1) the design, implementation, testing, and integration of an initial operating configuration of the NAS Processing System Network; 2) the systematic and evolutionary incorporation of advanced computer-system technologies to maintain a leading-edge performance cappability; and 3) the management and operation of the complex.

The new plan was presented to the various NASA Advisory Groups, the Office of Management and Budget, the Office of Science Technology and Policy and appropriate Congressional Subcommittees. It received strong support, and the Program was approved by Congress as a new start for NASA in the President's budget for fiscal year 1984. The Administrator of NASA at that time termed the NAS Program "the Centerpiece of NASA's Aeronautical Program."

Following Program approval, the development of the initial operating capability began in earnest. The in-house project team was expanded, and it was supplemented by a force of on-site contractor personnel. Procurements of both hardware and software were initiated and the evolving test-bed network was ready to receive the first High-Speed Processor, the Cray-2, in the Fall of 1985. After about 9 mo of test and integration, and with the help of a select group of users, the system was unveiled for national use in its Interim Initial Operating Configuration in July 1986. Within a few months the system was being used effectively by over 200 national users located both at Ames Research Center and at 20 remote sites.

The term "Interim Initial Operating Configuration" was selected to emphasize the fact that the system would not reach its first stage of maturity until it could be located in the new building that was being constructed as its ultimate home. Construction of this new building started in the Spring of 1985, and it was ready for occupancy at the end of 1986. The system was shut down for several weeks, dismantled, reassembled in the new building, and brought back into operation prior to meeting the goals of the Initial Operating Configuration. This conference celebrates the achievement of the goals of the Initial Operating Configuration, and commemorates the dedication of this new national capability. Plans are now well along for expanding the system and installing the second high-speed processor prior to reaching the goals of the first Extended Operating Configuration in 1988.

SUMMARY AND CONCLUDING REMARKS

A major milestone in aerodynamics research and development was reached in 1970 when, for the first time, computers began to solve problems not previously amenable to solution. Within several years, it became apparent that insufficient computer power would impose serious limitations on the growth of computational aerodynamics as a useful discipline. It was possible to calculate the flows about three-dimensional shapes such as wings and simple wing bodies, but only with highly approximate forms of the governing equations that neglected full treatment of important nonlinear and viscous phenomena. Consideration of more comprehensive physics forced the analyses to be restricted to simple two-dimensional shapes, such as airfoils or axisymmetric aircraft components. Even in this primative state, computational aerodynamics was recognized to have the potential to become a major complement to wind-tunnel testing. Working together, computers and wind tunnels would provide a formidable capability for designing aerospace vehicles.

Recognizing the potential importance of computational methods to the aerodynamics design process, a group of people at the Ames Research Center initiated an effort in 1975 to drive the development of a computer system powerful enough to take the next major step in the development and use of computational aerodynamics. This small initial effort grew with time and, in the fall of 1983, it became a major new program for NASA with two principal objectives: 1) to provide a supercomputer facility for the national aerospace community that would be maintained as close to the state of the art as possible, and 2) to serve as a pathfinder for the development and use of future supercomputer systems. The NAS Program will reach its first major milestone in March of 1987 when its initial capability was declared operational. Already, it was serving over 200 users nationwide, and plans were well underway for its extended operating capability having two powerful supercomputers, all of the necessary supporting equipment and well-established communications links.

Computational aerodynamics was in a relatively immature stage when the NAS Program was conceived in 1975. Even so, initial forecasts of the importance of the discipline to the country's aeronautics program and of the amount of computer power required to reach various plateaus have been remarkably accurate. Nothing has transpired in the intervening 12 yr that would temper the desire to push the developement of large-scale computer systems for the country's aerospace program as fast as the technology will allow. In fact,

supercomputers are now recognized as being absolutely essential for many fields of science and engineering, and all are benefiting from the efforts of the NAS Program to develop and maintain a leading-edge computational system.

REFERENCES

1. Magnus, R.; and Yoshihara, H.: Inviscid Transonic Flow Over Airfoils. AIAA J., Vol. 8, No. 12, Dec. 1970, pp. 2157-2162.

2. Ballhaus, W. F., Jr.; Magnus, R.; and Yoshihara, H.: Some Examples of Unsteady Transonic Flows Over Airfoils. Unsteady Aerodynamics, Vol. II, University of Arizona Press, 1975, pp. 769-791.

3. Borland, C. J.; and Rizzetta, D. P.: Nonlinear Transonic Flutter Analysis. AIAA Paper 81-0608-CP, May 1981.

4. MacCormack, R. W.: Numerical Solutions of the Interaction of a Shock Wave With a Laminar Boundary Layer. Lecture Notes in Physics, Vol. 8, Springer-Verlag, 1971, pp. 151-163.

5. Deiwert, G. S.: Numerical Simulation of High Reynolds Number Transonic Flow. AIAA Paper 74-603, June 1974.

6. Mansour, N. N.: Numerical Simulation of the Tip Vortex Off a Low-Aspect-Ratio Wing at Transonic Speed. NASA TM 85932, April 1984.

7. Flores, J.; Reznick, S. G.; Holst, T. L.; and Gundy, K.: Transonic Navier-Stokes Solutions for a Fighter-Like Configuration. AIAA Paper No. 87-0032, Jan. 1987.

8. Chapman, Dean R.: Computational Aerodynamics Development and Outlook. AIAA J. Vol. 17, No. 12, Dec. 1979, pp. 1293-1313.

9. Peterson, Victor L.: Impact of Computers on Aerodynamics Research and Development. IEEE Proc., Vol. 72, pp 68-79, Jan. 1984.

10. Inouye, M. (ed.): Future Computer Requirements for Computational Aerodynamics. NASA CP 2032, 1978.

Table 1.- Governing equations, results, and computer requirements for computational aerodynamics.

APPROXIMATION	CAPABILITY	GRID POINTS REQUIRED	COMPUTER REQUIREMENT
LINEARIZED INVISCID	SUBSONIC/SUPERSONIC PRESSURE LOADS VORTEX DRAG	3×10^3 PANELS	1/10 CLASS VI
NONLINEAR INVISCID	ABOVE PLUS: TRANSONIC PRESSURE LOADS WAVE DRAG	10^5	CLASS VI
REYNOLDS AVERAGED NAVIER-STOKES	ABOVE PLUS: SEPARATION/REATTACHMENT STALL/BUFFET/FLUTTER TOTAL DRAG	10^7	30 X CLASS VI
LARGE EDDY SIMULATION	ABOVE PLUS: TURBULENCE STRUCTURE AERODYNAMIC NOISE	10^9	3000 X CLASS VI
FULL NAVIER-STOKES	ABOVE PLUS: LAMINAR/TURBULENT TRANSITION TURBULENCE DISSIPATION	10^{12} TO 10^{15}	3 MILLION TO 3 BILLION CLASS VI

Table 2.- Historical role of the Government as a prime driver in advancing computer capability.

TIME	DRIVING NEED	SPONSOR	COMPUTER DEVELOPED	KEY TECHNOLOGY	COMMERCIAL FOLLOW-ONS
MID 1940'S (WW II)	MULTITUDE OF BALLISTIC TABLES	ORL	ENIAC	VACUUM TUBE ELECTRONIC COMPUTING	IBM 701, UNIVAC I
EARLY-MID 1950'S	DEW AIR DEFENSE FOR TRACKING COMBER FLEET	USAF	AN FSQ-7	MAGNETIC CORE MEMORY	IBM 700
EARLY 1950'S	SUPERIOR DESIGN CAPABILITY FOR SMALL NUCLEAR DEVICES	AEC	CDC G300	INTEGRATED CIRCUITS	CDC 7000, IBM 370
LATE 1950'S	ANTI-ICBM CONTROL SYSTEM INEED ELIMINATED POLITICALLY PRIOR TO COMPLETION IN 1972)	DARPA	ILLIAC IV	SEMICONDUCTOR MEMORY AND PARALLEL PROCESSING	CDC STAR, CRAY-1*
CIRCA 1990	SUPERIOR DESIGN CAPABILITY FOR AIRCRAFT	NASA	NAS PROCESSING SYSTEM NETWORK	NETWORKING OF SUPERCOMPUTERS COMMON USER INTERFACE	

Table 3.- NAS User Interface Group.

FUNCTION

- INFORMATION CHANNEL BETWEEN USER COMMUNITY AND PROJECT
- IDENTIFY AND DISCUSS USER-ORIENTED ISSUES, e.g., REMOTE ACCESS

PARTICIPATING ORGANIZATIONS

- AIRFRAME COMPANIES
  BOEING AEROSPACE, GENERAL DYNAMICS, GRUMMAN AEROSPACE, LOCKHEED-CALIF., LOCKHEED-GA., McDONNELL DOUGLAS, NORTHROP, ROCKWELL, VOUGHT
- ENGINE COMPANIES
  DETROIT DIESEL ALLISON, GENERAL ELECTRIC, PRATT AND WHITNEY
- DEFENSE DEPARTMENT
  AFWAL, AEDC, BRL, DTNSRDC, NUSC
- GENERAL AVIATION
  GENERAL AVIATION MANUFACTURERS ASSOC. (GATES-LEARJET)
- ROTORCRAFT
  AMERICAN HELICOPTER SOCIETY (UNITED TECHNOLOGY CORP. RES. CENTER)
- UNIVERSITIES
  STANFORD, UNIVERSITY OF COLORADO, SCRIPPS INSTITUTION OF OCEANOGRAPHY, PRINCETON, MASSACHUSSETTS INSTITUTE OF TECHNOLOGY
- NATIONAL SCIENCE FOUNDATION (NSF)
- NATIONAL CENTER FOR ATMOSPHERIC RESEARCH (NCAR)
- NASA
  AMES, GODDARD, LANGLEY, LEWIS

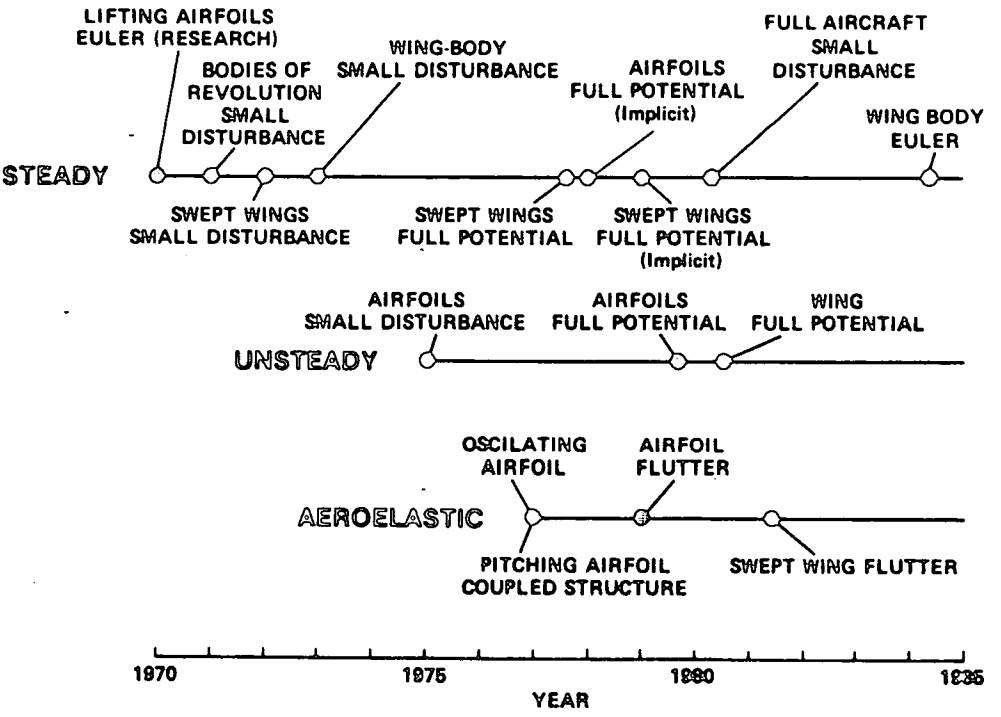


Figure 1.- Milestones in the development of computational aerodynamics; inviscid transonic flows.

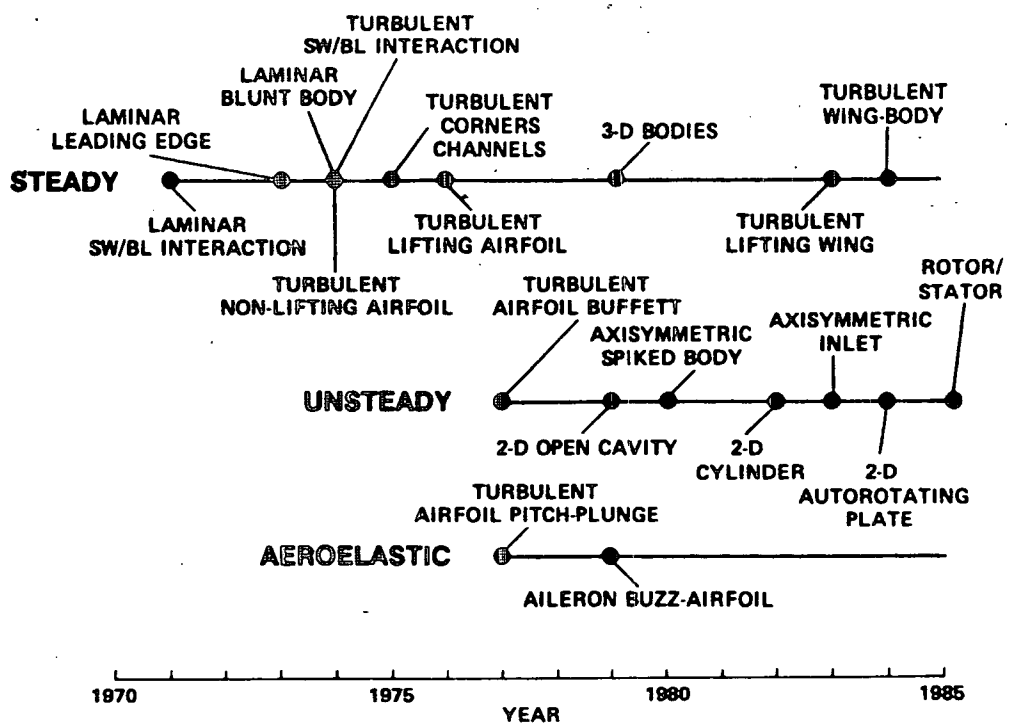


Figure 2.- Milestones in the development of computational aerodynamics; compressible viscous flows.

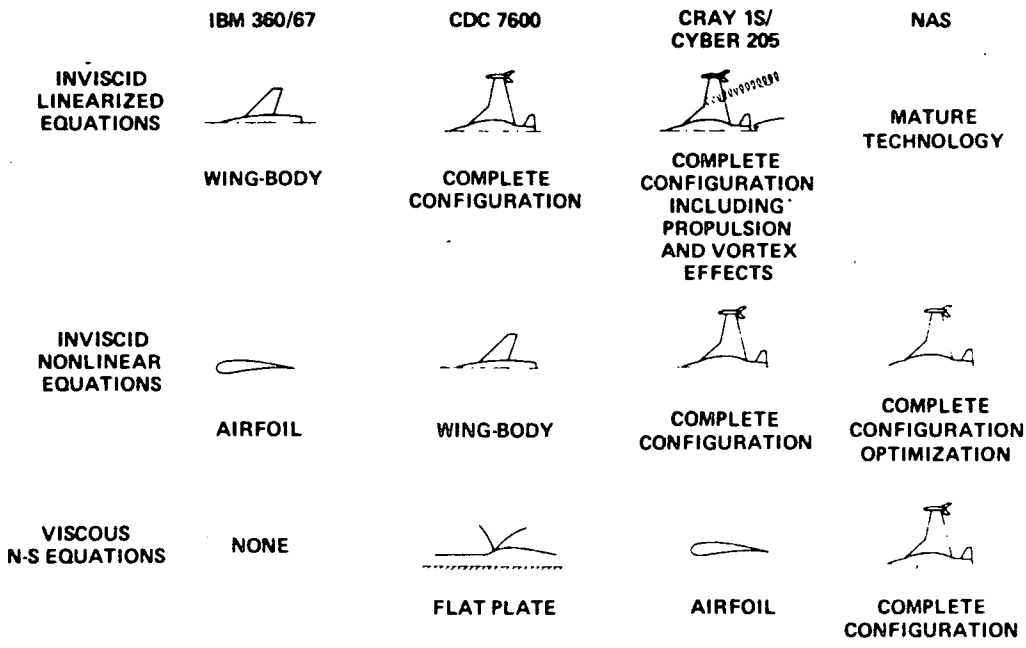


Figure 3.- Pictorial representation of the effect that increasing computer power has had on computational aerodynamics.

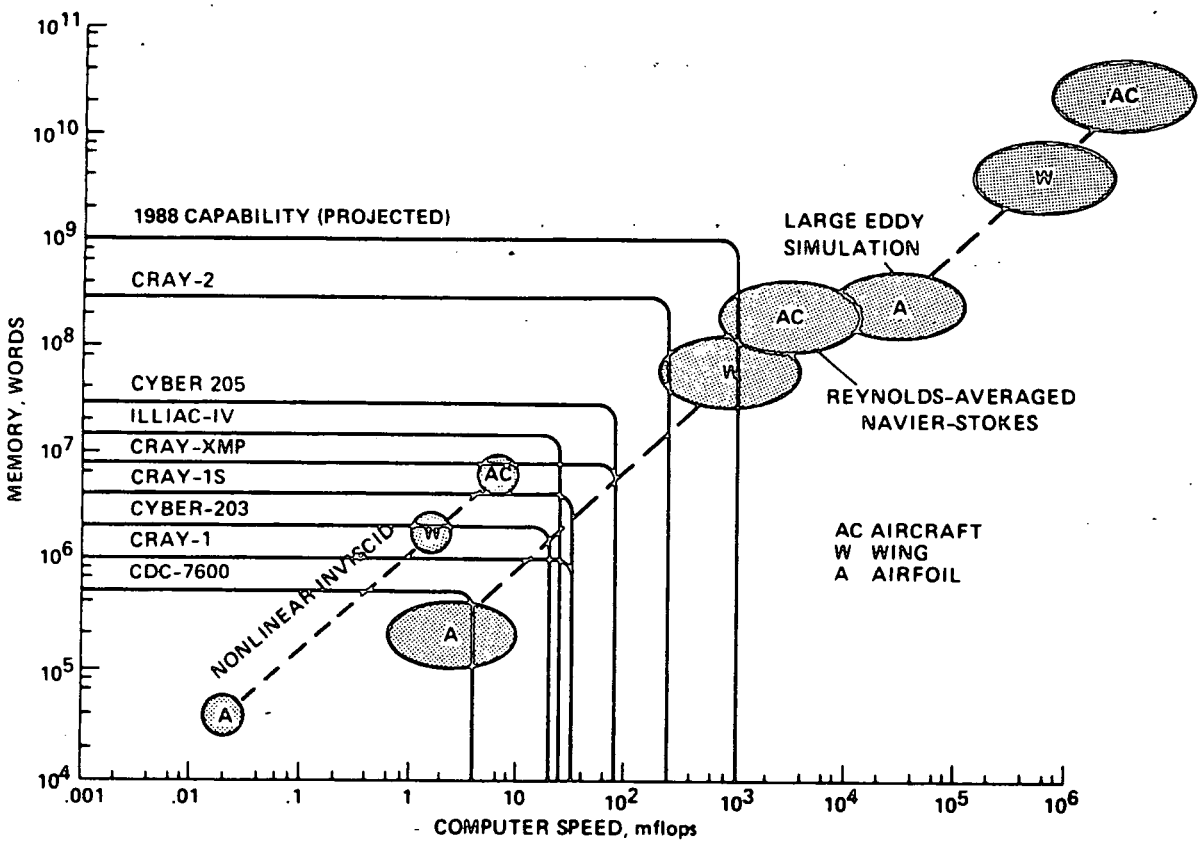


Figure 4.- Computer speed and memory requirements for aerodynamic calculations compared with the capabilities of various machines; 15-min runs with 1985 algorithms.

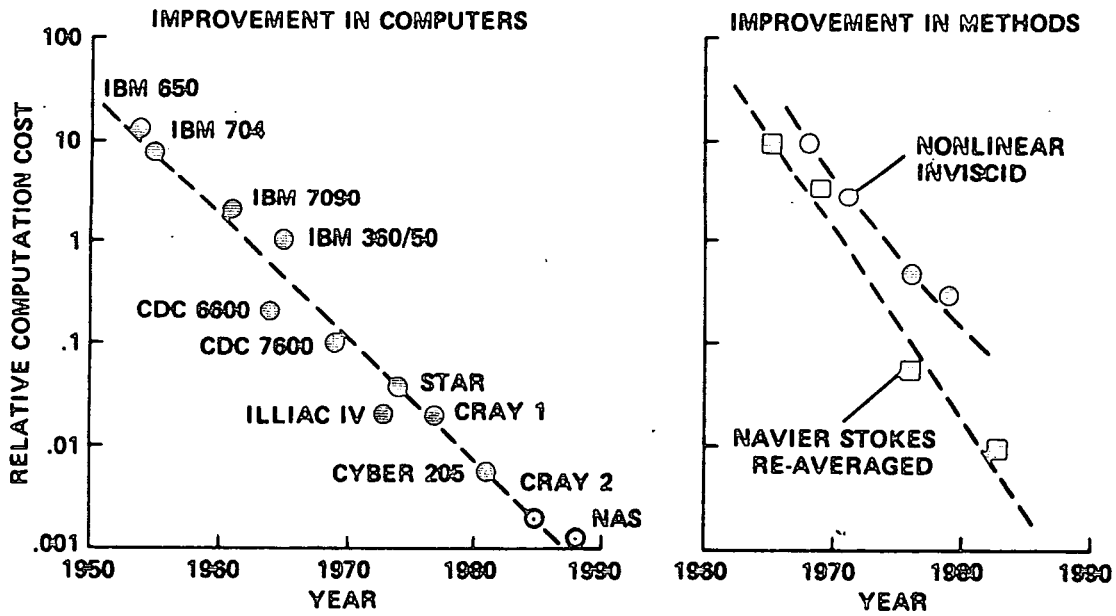


Figure 5.- Comparison of numerical simulation cost trend resulting from improvements in computers with that resulting from improvements in algorithms.

The Federal High Performance Computing Program

Executive Office of the President
Office of Science and Technology Policy
September 8, 1989

THE FEDERAL HIGH PERFORMANCE COMPUTING PROGRAM

High Performance Computing Systems

- Research for Future Generations
- System Design Tools
- Advanced Prototype Development
- Evaluation of Early Systems

Advanced Software Technology and Algorithms

- Support for Grand Challenges
- Software Components and Tools
- Computational Techniques
- High Performance Computing Research Centers

National Research and Education Network

- Interagency Interim NREN
- Gigabits Research and Development
- Deployment of Gigabits NREN
- Structured Transition to Commercial Service

Basic Research and Human Resources

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20506

In November 1987, my predecessor, William R. Graham, transmitted to Congress A Research and Development Strategy for High Performance Computing. That report laid out a five-year strategy for federally supported R&D on high performance computing, including hardware for state-of-the-art supercomputers, software, computer networks, and supporting infrastructure. It was written with the assistance of the Committee on Computer Research and Applications under the OSTP Federal Coordinating Council for Science, Engineering, and Technology (FCCSET). This strategy document was to be followed by a detailed program plan.

I am pleased to transmit to Congress that program plan -- the result of an intense interagency effort by a special task force within the Committee on Computer Research and Applications. Following the general organizational structure of the 1987 strategy report, it lays out a broad R&D policy and program plan designed to advance U.S. leadership in high performance computing. This plan calls for a federally coordinated government, industry, and university collaboration to accelerate the development of high speed computer networks and to accelerate the rate at which high performance computing technologies -- both hardware and software -- can be developed, commercialized, and applied to leading-edge problems of national significance.

High performance computing is a vital and strategic technology, exerting strong leverage on the rest of the computer industry and other cutting-edge areas. However, U.S. leadership and diversity in the supercomputer industry itself has declined dramatically; and history shows that a scant 15 years separates the first appearance of a top-of-the-line supercomputer from the appearance of that same computing power in the higher end of the personal computer market. A future national high speed computer network could have the kind of catalytic effect on our society, industries, and universities that the telephone system has had during the twentieth century.

We cannot afford to cede our historical leadership in high performance computing and in its applications. We need to encourage the dynamism of the U.S. computer industry and, hence, our economy. I would ask all of the federal agencies with research programs in high performance computing to work toward implementing the recommendations in this report.

D. Allan Bromley
Director

Foreword

High Performance Computing is a powerful tool to increase productivity in industrial design and manufacturing, scientific research, communications, and information management. It represents the leading edge of a multi-billion dollar world market, in which the U.S. is increasingly being challenged. A strong, fully competitive domestic high performance computer industry contributes to U.S. leadership in critical national security areas and in broad sectors of the civilian economy, including the technical base for many national economic and military security needs. For this reason we are initiating the preliminary planning to address this important U.S. technology.

GOALS

Accordingly, the goals of the Federal High Performance Computing (HPC) Program are to:

- Maintain and extend U.S. leadership in high performance computing, and encourage U.S. sources of production;
- Encourage innovation in high performance computing technologies by increasing their diffusion and assimilation into the U.S. science and engineering communities; and
- Support U.S. economic competitiveness and productivity through greater utilization of networked high performance computing in analysis, design, and manufacturing.

COMPONENTS

The HPC Program is implemented through four complementary, closely coordinated, multidisciplinary Components:

- High Performance Computing Systems;
- Advanced Software Technology and Algorithms;
- The National Research and Education Network; and
- Basic Research and Human Resources.

POLICY

The Federal High Performance Computing (HPC) Program features increased cooperation between business, academia and government. While each of these sectors will retain its current role, the success of this Program will depend in large part on an effective transition from R&D to commercialization—an outcome of successful cooperation among the above sectors.

The measure of success of this Program in the area of R&D will be an increased rate of development of new computing concepts, systems, and architectures. A longer term measure of success will be the rate at which this technological progress shows up in commercialized products. The HPC Program will be consistent with the traditional roles of government, business and academia.

Specifically:

- The government will provide R&D support for HPC and will coordinate R&D among its agencies;
- Business will be the decision maker and source of capital investment for commercialization of HPC technology in response to its assessment of market opportunities; and
- Universities and Federal laboratories will be the primary institutions receiving government funding under this Program.

The government will, in addition, foster a number of mechanisms for increased collaboration and interaction among government, business and universities. Specifically:

- The government will continue to serve as a market for commercial prototypes and for commercial products. This particularly will be the case in the defense sector. These markets will exist in U.S. laboratories, Federal agencies, university centers of excellence and industrially led consortia;
- The government will assist in the development of industrially-led consortia in cases where appropriate (an existing example is SEMATECH); and
- The government will promote centers of excellence, jointly funded and staffed by government, academia, and industry. Technology transfer to industry from government and academia will happen automatically as a result of this ongoing collaboration.

Foreign policy objectives will be supported through existing or future international science and technology agreements. "Symmetry and reciprocity," protection of U.S. proprietary interests, and enforcement of intellectual property rights will continue to be guiding principles.

The Federal High Performance Computing Program will ensure the broadest possible national benefit by addressing:

- Many problems susceptible to computational solution;
- A wide geographic and demographic distribution; and
- The inclusion of government, academia and industry.

STRATEGY

To achieve the policy goals of the HPC Program, our strategy is to:

- Support computational advances through R&D effort to address U.S. scientific and technical challenges;
- Reduce the uncertainties to industry for development and use of this technology though increased cooperation among government, industry and academia and the continued use of government and government-funded facilities as a market for HPC prototypes and commercial products;

- Support the underlying research, network and computational infrastructures on which U.S. high performance computing technology is based; and
- Support the U.S. human resource base to meet needs of industry, academia and government.

ROLE OF FEDERAL, ACADEMIC AND INDUSTRIAL SECTORS

Federal agencies

- Funding for the Program will come from agencies their annual appropriations;
- User agencies will continue to define their respective missions and goals, though guided by the High Performance Computing Program goals and objectives; and
- OSTP, through its Federal Coordinating Council on Science, Engineering, and Technology (FCCSET) Committee on Computer Research and Applications, will assist the agencies as part of its continuing responsibility for coordination and policy guidance. OSTP will also assist by recommending special computational opportunities. However, final priority setting will reside with the respective agencies.

Academia

Universities and colleges will participate in the HPC Program in the following ways:

- Responding to agency program announcements;
- Forming consortia with government and industry;
- Focusing research capabilities on specific areas of computational science;
- Enhancing curricula to take advantage of new generations of computing technologies, attracting additional manpower into various disciplines of computational science; and
- Bringing the Program to the attention of State leaders for potential leveraging of Federal funds.

Industry

- Private industry will develop hardware, software, and networks in response to the Program. Commercialization will be at the initiative and discretion of private industry;
- Industry will join and help finance university or government laboratory R&D activities (at its choosing) to obtain access to expertise and government funded facilities. As a result of these collaborative relationships, the partnership will supply industry with R&D and technology information;
- A broadly representative industry body will assist in making long-range demand and robustness projections for: high capacity research networks; the spectrum of

computer architectures; the adequacy of software development; and the level of the manpower pool. This body will help assure a smooth transition between successive generations of high performance computing systems; and

- Private industry suppliers will provide the network services to Federal agencies in the first two stages of the National Research and Education Network. Industry should plan to operate the NREN fully as soon as feasible.

FUNDING OF THE HPC PROGRAM

The magnitude of the program envisioned by this Program will require major new Federal R&D investment. It is assumed that existing Federal base funding for computer and information science and technology research and development, roughly $500 million annually, will continue. Preliminary planning estimates suggest that the first year of the program would require an augmentation of $150 million, which would then grow to an incremented annual level of $600 million by the fifth year.

MANAGEMENT OF THE HPC PROGRAM

The components of the Program will be managed by existing Federal agencies.

Oversight of the HPC Program will be the responsibility of the Office of Science and Technology Policy with the assistance of the FCCSET Committee on Computer Research and Applications and the help of a High Performance Computing advisory panel which will report to the Director of OSTP:

- The HPC advisory panel will interact regularly with the FCCSET Committee on Computer Research and Applications; and
- The HPC advisory panel will have representation from all sectors and will monitor the progress of the Program for cross-sector balance, breadth of applicability, network security, competitiveness versus international cooperation, and technology transfer effectiveness.

SCOPE OF THIS REPORT

This report is designed for agency-level planning purposes and does not represent the Administration's approval or support of any program not included in the President's budget requests. Programs discussed in this document are subject to budget constraints and Administration approval.

TABLE OF CONTENTS

1. Executive Summary

*High Performance Computing** is a pervasive and powerful technology for industrial design and manufacturing, scientific research, communications, and information management. A strong U.S. high performance computer industry contributes to our leadership in critical national security areas and competitiveness in broad sectors of the civilian economy.

The goals of the High Performance Computing Program are to:

Goals

- Maintain and extend U.S. leadership in high performance computing, and encourage U.S. sources of production;
- Encourage the pace of innovation in high performance computing technologies by increasing their diffusion and assimilation into the U.S. science and engineering communities; and
- Support U.S. economic competitiveness and productivity through greater utilization of networked high performance computing in analysis, design, and manufacturing.

Strategy

These goals will be accomplished through Federally coordinated government, industry, and university collaboration to:

- Support computational advances through a more vigorous R&D effort to expedite solutions to U.S. scientific and technical challenges;
- Reduce the uncertainties to industry for R&D and use of this technology through increased cooperation between government, industry and academia and the continued use of government and government-funded facilities as a market for HPC early commercial products;
- Support the underlying research, network and computational infrastructures on which U.S. high performance computing technology is based; and
- Support the U.S. human resource base to meet needs of industry, academia and government.

* *High performance computing* refers to the full range of advanced computing technologies including existing supercomputer systems, special purpose and experimental systems, and the new generation of large scale parallel systems.

The HPC Program

The Program will consist of four complementary, coordinated components in each of the key areas of high performance computing. The components are planned carefully to produce not only long term results but a succession of intermediate national benefits. Figure 1 shows the relationship of the components of the Program. The High Performance Computing Program will build on those programs already in place, providing additional funds in carefully selected areas to meet its goals. Selected computational challenges, which will have significant effect on national leadership in science and technology, will be used as focal points for these efforts.

High Performance Computing Systems: The United States' leadership in supercomputing is increasingly being challenged. We have developed new, more powerful supercomputing architectures based on innovations. Particularly in parallel processing, we must capitalize on these innovations. To do this, a long range effort involving Federal support will be required for basic research on high performance computing technology and the appropriate transfer of research and technology to U.S. industry, consisting of efforts in the following areas:

- Research for future generations of computing;
- System design tools;
- Advanced prototype development; and
- Evaluation of early systems.

Advanced Software Technology and Algorithms: Historically, software improvements have increased computational performance much more than hardware investments. Yet software productivity is generally poor, and existing software can seldom be re-used without modification. In computing systems for industrial, scientific and military applications, software costs have exceeded those of hardware more than fivefold. Advances in software will be critical to the success of high performance computers with massively parallel architectures. To improve software productivity, an interagency effort will support joint research among government, industry and universities to improve basic software tools, data management, languages, algorithms, and associated computational theory with broad applicability for the *Grand Challenge** problems. These complex problems will require advances in software that have widespread applicability to computational problems in science and technology.

* A *Grand Challenge* is a fundamental problem in science or engineering, with broad economic and scientific impact, that could be advanced by applying high performance computing resources.

1. Executive Summary

Effort in this component focuses on:

- Support for Grand Challenges;
- Software components and tools;
- Computational techniques; and
- High performance computing research centers.

National Research and Education Network: For the past decade technology developed by the U.S. has been available to eliminate distance as a factor in computer access and in collaborations among high technology workers. To maintain our leadership, the U.S. government, together with industry and universities, will jointly develop a high-speed research network to provide a distributed computing capability linking government, industry and higher education communities. This network will serve as a prototype for future commercial networks which will become the basis for a distributed industrial base. This component will consist of:

- An interagency effort to establish an interim National Research and Education Network;
- Research and development for a billions of bits per second (gigabits) network adequate to support national research needs;
- Deployment of the gigabits National Research and Education Network; and
- Structured transition to commercial service.

Basic Research and Human Resources: U.S. universities are not meeting the expanding needs of industry for trained workers in computer technology. There is not an adequate number of high quality computer science departments in this country, and many industrial and Federal laboratories have inadequate research capabilities. Furthermore, existing university, government, and industrial groups do not collaborate effectively enough, and their interdisciplinary activities are too limited. To correct these deficiencies a long term effort to support basic research in computer science and engineering (creating computing systems) will be established by building upon existing programs. This component will also establish industry, university, and government partnerships to improve the training and utilization of personnel and to expand the base of research and development personnel in computational science and technology (using computers).

Organization

Leadership of the Program is the responsibility of the Office of Science and Technology Policy, through the Federal Coordinating Council on Science, Engineering and Technology (FCCSET) Committee on Computer Research and Applications, whose members include representatives of the key agencies that fund R&D in high performance computing. Duties and responsibilities of the Committee include:

- Interagency planning and coordination;
- Technology assessment;
- Policy recommendations to OSTP; and
- Formal annual reports of progress to OSTP.

A High Performance Computing Advisory Panel will be formed, consisting of eminent individuals from government, industry, and academia. Members of the Advisory Panel will be selected by and will report to the Director of OSTP. The Panel will provide the Director and the Committee on Computer Research and Applications with an independent assessment of:

- Progress of the Program in accomplishing its objectives;
- Continued relevance of the Program goals over time;
- Overall balance among the Program Components; and
- Success in strengthening U.S. leadership in high performance computing, and integration of these technologies into the mainstream of U.S. science and industry.

This implementation plan was prepared by the FCCSET Committee on Computer Research and Applications under the leadership of the Office of Science and Technology Policy. It represents a broad spectrum of government, industrial and university interests. The Committee has established subcommittees that will be responsible for planning, organizing, monitoring and coordinating the components of the Program.

SCOPE OF THIS REPORT

This report is designed for agency-level planning purposes and does not represent the Administration's approval or support of any program not included in the President's budget request. Programs discussed in this document are subject to budget constraints and Administration approval.

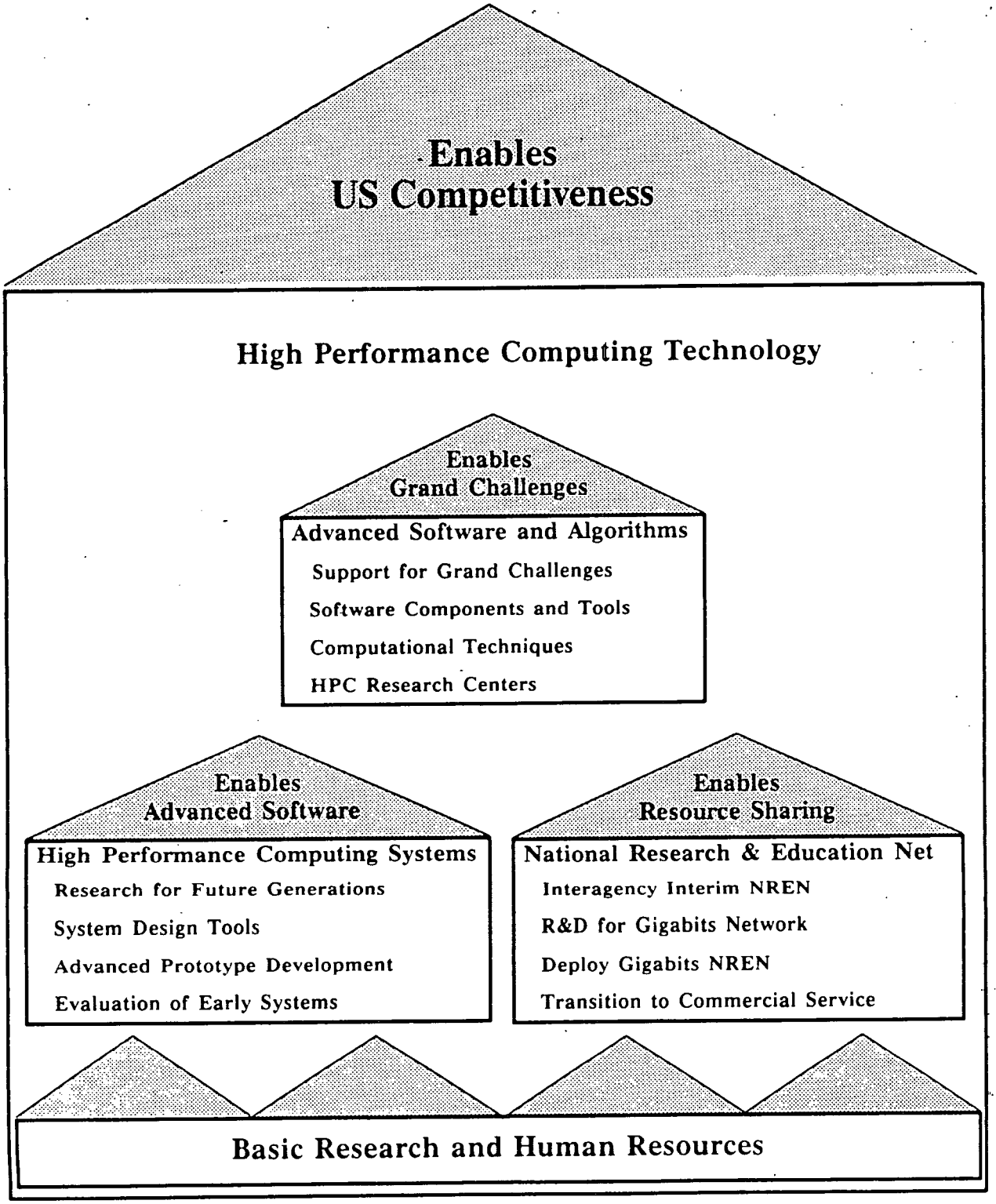


Fig. 1 - Relationship of HPC Program Components

2. Introduction

Purpose and Scope of Report

The purpose of this document is to provide the initial implementation plan for the U.S. High Performance Computing Program. This plan encompasses the first five year period and provides for periodic reviews to be conducted by the Federal Coordinating Council on Science, Engineering and Technology (FCCSET) with the participation of government, industry, and university representatives.

This document discusses:

- National economic and technical issues associated with high performance computing;
- Goals and strategy of this Program;
- Plans for synergistic government, industrial, and university participation;
- Organizational structure to coordinate, manage and review the Program program;
- Economic and technical benefits of the Program; and
- Proposed budget for the first five years of the Program.

This implementation plan was prepared by the FCCSET Committee on Computer Research and Applications under the leadership of the Office of Science and Technology Policy. It represents a broad spectrum of government, industrial, and university interests. This report is designed for agency-level planning purposes and does not represent the Administration's approval or support of any program not included in the President's budget request. Programs discussed in this document are subject to budget constraints and Administration approval.

Background

The Federal Coordinating Council on Science, Engineering and Technology (FCCSET), chartered by the Office of Science and Technology Policy (OSTP), coordinates Federal interagency activities of broad national interest. The FCCSET Committee on Computer Research and Applications serves as the forum for developing a national agenda for computing technology needs, opportunities, and trends.

This FCCSET Committee has examined the scientific, technological and economic effects of high performance computing. The Committee issued reports as early as 1983

that assessed the status of high performance computing and possible supporting government activities. These studies have consistently demonstrated the need for a strategy to coordinate high performance computing related activities in the government, industrial and university sectors. Dramatic increases in foreign investments in computer related technology have been noted, which challenge the world leadership of the U.S. computing industry. The studies also emphasized that advances in critical areas of national security and broad sectors of the civilian economy depend strongly on high performance computing technology.

The unprecedented power of high performance computing systems has created a new mode of scientific research: computational investigations that complement the traditional modes of experiment and theory. Computational research is being applied to a wide range of scientific and engineering problems called Grand Challenges. *A Grand Challenge is 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.* While the Grand Challenges are already being addressed to some extent using existing supercomputers, progress is often severely limited by current computer speeds and memory capacities. Examples of Grand Challenges are:

(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 our understanding of the fundamental nature of matter, including quantum chromodynamics and condensed matter theory.

(6) Machine vision to enable real-time analysis of complex images for control of mechanical systems.

The sample Grand Challenge areas provided in Appendix A are representative of the science and technology areas that will be affected by application of leading edge computational resources and supporting systems. Figure 2 illustrates some of the Grand Challenges that can be adequately addressed through existing high performance computing technology and problems that could be attacked much more successfully with a thousandfold increase in performance.

2. Introduction

Agency Activities: In the early 1980's, Federal agencies initiated programs that provide the basis for the opportunities described in this Program. The NSF established the National Supercomputer Centers to provide high performance computers to the science and engineering community and interconnected them with the research community via the NSFNET. The centers and network have stimulated the development of innovative computational approaches to a wide range of scientific and engineering problems related to the Grand Challenges. NSF also reorganized to create a new Directorate of Computer and Information Science and Engineering (CISE) with increased emphasis and funding for computer and computational disciplines, with a focus on computer networking as a tool for scientific and engineering research.

DARPA initiated the Strategic Computing program to accelerate development of an alternate approach to building high performance computer systems. This program focuses on large scale parallel systems, custom VLSI and associated software, including symbolic processing for the advanced functionality characterized by artificial intelligence. Strategic Computing stimulated the first generation of commercially available scalable parallel computer systems using conservative components and packaging. Early production models of these systems were acquired by several agencies for experimental use. A second generation of these systems is being developed, using custom VLSI. The military services have participated in this program, providing applications focus and technical consultation.

The Office of Naval Research, Air Force Office of Scientific Research, and Army Research Office have separately sponsored important research and development in basic research for advanced computing.

The DOE expanded the National Magnetic Fusion Computer Center and its MFE Network to serve all energy research users in national laboratories, universities, and industry. Several of the National Laboratories have formed computational groups to experiment with novel high performance computers and to develop algorithms that exploit the power of those computers. Special funding was provided to enable university, industry, laboratory collaborations with the national laboratories to acquire parallel computer prototypes to test ideas for advanced high performance computing architectures.

NASA upgraded the computational capability at several of its research and flight centers and established a data network to link them together. At the Ames Research Center, the Numerical Aerodynamics Simulation (NAS) was set up to provide a focused attack on computational aerodynamics employing the highest powered computers available surrounded by data reduction and visualization systems.

HPC Strategy: In 1986 Congress requested that OSTP conduct a study of the critical problems and options for communication networks that support the U.S. high

performance computing environment. The charter of the FCCSET Committee on Computer Research and Applications was broadened to include the technical aspects of this study. A number of working groups were formed to ensure a perspective that spanned all aspects of the U.S. high performance computing environment. In addition a consortium of government, industry and university experts focused on national infrastructure requirements for high performance computing.* The FCCSET study is documented in "A Research and Development Strategy for High Performance Computing" also known as the *High Performance Computing Strategy (HPC Strategy)*, published by the Office of Science and Technology Policy (included as Appendix B). It provides the foundation for this Program.

The *HPC Strategy* findings were:

- A strong domestic high performance computer industry contributes to maintaining U.S. leadership in critical national security areas and in broad sectors of the civilian economy.
- Research progress and technology transfer in software and applications must keep pace with advances in computing architectures and microelectronics.
- The U.S. faces serious challenges in networking technology which could become a barrier to the advance and use of computing technology in science and engineering.
- Federal research and development funding has established laboratories in universities, industry, and government which have become the major sources of innovation in the development and use of computing technology.

The recommendations of the *HPC Strategy* form the basis for the four components of the High Performance Computing Program.

Four National Research Council reports issued in the period following publication of the *HPC Strategy* have confirmed its findings and emphasized the need to carry out its recommendations: *Toward a National Research Network* (1988), *The National Challenge in Computer Science and Technology* (1988), *Global Trends in Computer Technology and Their Impact on Export Control* (1988), and *Information Technology and the Conduct of Research* (1989).

In December 1988, the Office of Science and Technology Policy charged the FCCSET Committee on Computer Research and Applications to develop this implementation plan for the High Performance Computing Program. The goals, strategy, and actions to implement the Program are discussed in the following sections.

* *A National Computing Initiative*,
Society for Industrial and Applied Mathematics, Philadelphia, PA 1987.

What is High Performance Computing?

High Performance Computing refers to a productive computing environment that includes high performance components, system and applications software, networking, and the underlying research and human resource infrastructure.

High performance computing systems are those at the forefront of the computing field in terms of computational power, storage capability, input/output bandwidth, and software. These systems include high speed vector and pipeline machines, special purpose and experimental systems, scalable parallel architectures, and associated mass storage systems, input/output units, and systems software. Underlying these advanced systems are microelectronics, optoelectronics, logic devices, and storage technologies.

Advanced software technology and algorithms includes general-purpose operating systems for high performance computer systems and tools and utilities, such as compilers, analysis tools, debuggers, and data management systems. Mathematical algorithms and other general purpose libraries facilitate the use of high performance computers for science and engineering. Software tools will allow high performance computing systems to be embedded transparently in a distributed environment which includes applications specific software and other specialized methods and algorithms. The technology base required to build such environments includes software engineering and data management tools, and basic research in high-level languages and algorithms. Improving these capabilities will greatly enhance scientific and engineering software productivity.

Computer network technology consists of communications and switching capable of providing a very high speed backbone on which the high performance computing environment is distributed. Internetworking and feeder network technology connects local or mid-level high speed networks to the national high speed network. User services such as directories are also essential components of an effective networking environment. Advanced networks will provide improved access to high performance computing systems and increased collaboration opportunities for universities, industry, and government.

Development of high performance computing environments requires a long term, continuing investment in basic research over a wide spectrum of computer and computational science and engineering. A basic infrastructure of knowledge, research, computing facilities, and people are required to create and exploit high performance computing technology.

Why is High Performance Computing Important to the U.S.?

During the last two decades computing has become an important complement to experimental and theoretical research. Computer aided design and engineering

techniques are replacing manual ones. Computer assisted and automated manufacturing is increasing productivity and improving the value and reliability of industrial products, while reducing the time required for engineering and manufacturing cycles. New knowledge and new industries are increasingly dependent upon computing.

Most of these advances in computing have originated in the United States. However, many of them have been most successfully applied in other countries, where their use has eroded the competitive edge that the U.S. had previously enjoyed. This Program is intended to maintain the U.S. edge by focusing our research advantage in high performance computing toward applications with high value to our economy and national security. Fortunately, current U.S. leadership in high performance computing offers a strategic opportunity to maintain our industrial momentum. The HPC Program provides a way to do this.

The national economic benefits of a strong high performance computing industry are recognized and pursued by other countries. Those nations have formed and funded collaborations between their private and public sectors. Their successes constitute vigorous competition for technological and economic leadership in high performance computing. Foreign computing industries benefit tremendously from government support. To retain our leadership, domestic industrial efforts must be encouraged by a strategy that shares the economic risk of innovation in this capital-intensive field.

National economy: High performance computing is by definition the leading edge of computing technology, which in turn supports many areas of science and technology. Computing constitutes a significant portion of the U.S. economy. For example, in 1988 the U.S. computing industry accounted for 10% of GNP, and almost 10% of all capital investment.* The pace of innovation that it sets pervades the domestic computing industry technology and economics. In terms of capability, today's supercomputer is tomorrow's desk-top workstation and the following day's classroom tool. Thus, U.S. competitive success in the world computing market is supported by leadership in high performance computing.

National security: High performance computing technology is used in critical national security areas. Examples include advanced computer systems architectures, computer network communication technology, and signal processing techniques. Continued acceleration of this technology, including availability of U.S. sources of production, is important to U.S. national security.

Science and technology: High performance computing provides a basis for other innovative scientific and engineering efforts. The pace of rapidly developing technologies, such as robotics, artificial intelligence, communications, high definition

* *The National Challenge in Computer Science and Technology*, National Academy Press, Washington, D.C., 1988, p. 7.

television (HDTV), campus network applications, semiconductor design, superconductivity, transportation, speech recognition, and data visualization are all dependent on a strong and innovative high performance computing industry.

Manufacturing: High performance computing constitutes an important tool for many industries. Its use in simulation and design improves the productivity of large industries such as aircraft production and automobile manufacturing and is rapidly being extended to other industries. Recent vigorous growth in use of high performance computing in electronics, energy, chemical and pharmaceutical industries illustrates the role of computing in the long term strength of the U.S. economy.

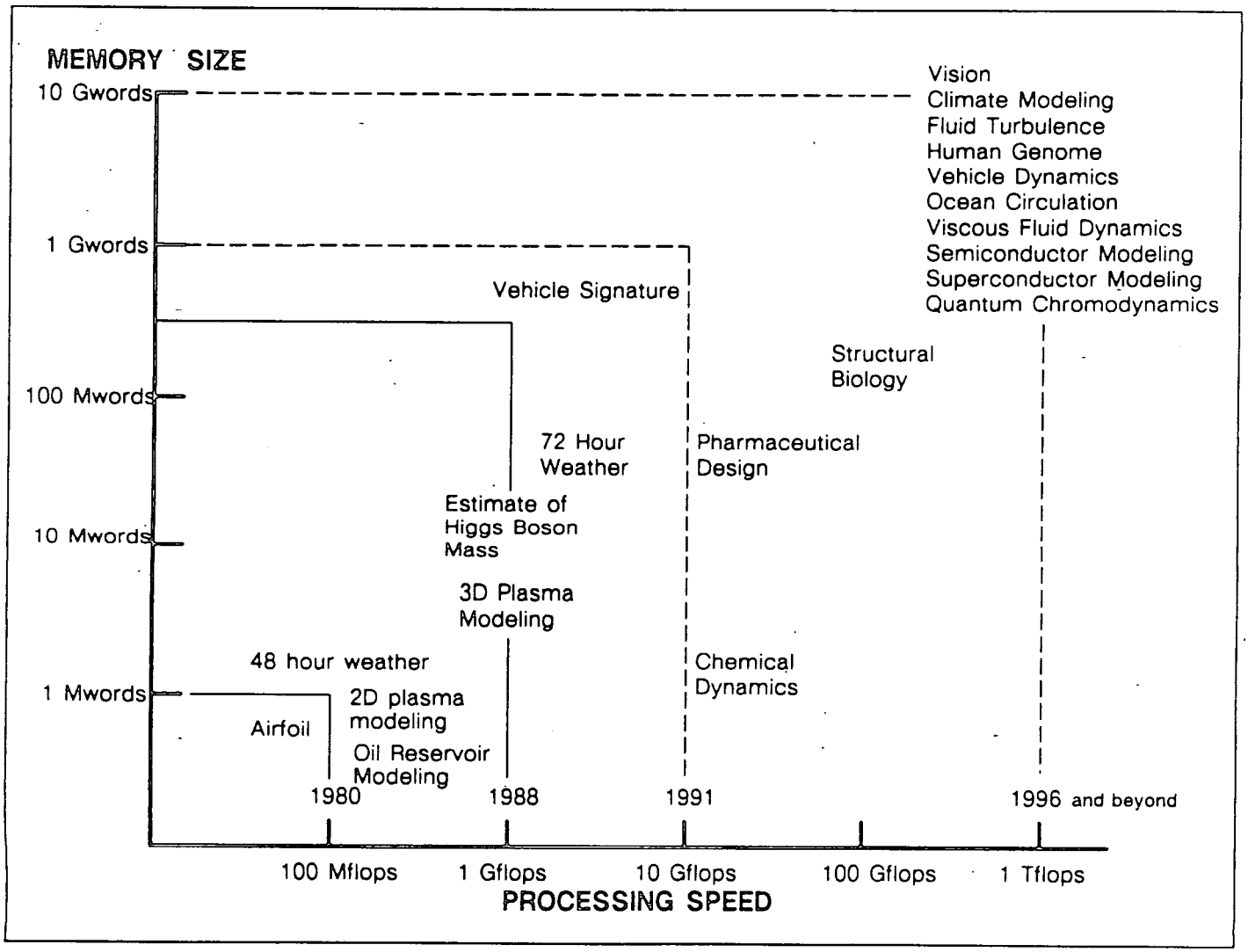


Fig. 2 – Some Grand Challenges and their Projected Computational Requirements

3. Program Plan

Introduction

The goals of the High Performance Computing Program are to:

- Maintain and extend U.S. leadership in high performance computing, and encourage U.S. sources of production;
- Encourage innovation in high performance computing technologies by increasing their diffusion and assimilation into the U.S. science and engineering communities; and
- Support U.S. economic competitiveness and productivity through greater utilization of networked high performance computing in analysis, design, and manufacturing.

To achieve these goals, a strategy has been established to:

- Support computational advances through R&D efforts to address U.S. scientific and technical challenges;
- Increase the use of this technology by reducing the uncertainties to industry for R&D and by increasing cooperation among government, industry, and academia;
- Continue use of government and government-funded facilities as a market for HPC early commercial products;
- Support the underlying research, network and computational infrastructures on which U.S. high performance computing technology is based; and
- Support the U.S. human resource base to meet needs of industry, academia and government.

The HPC Program is composed of four coordinated program components: High Performance Computing Systems, Advanced Software Technology and Algorithms, the National Research and Education Network, and Basic Research and Human Resources. Each of the four program components stimulates the development of progressively more advanced products for use throughout computing technology. The four areas build upon each other and upon the existing research base, as illustrated in Figure 1. Each component carries out a recommendation of the November 20, 1987 *HPC Strategy* (see Appendix C.)

Although the program components are described separately, they are interdependent, so that success of the Program depends on balanced support for all of them. For example, the development of high performance computing systems depends on development of advanced software technology and algorithms, because algorithm and software requirements largely determine the corresponding architecture of successful

computing systems. Similarly, as the new computing systems become available, new algorithms and software systems are required to take advantage of their capabilities and allow the systems to be used in practical ways.

The High Performance Computing Program requires an unprecedented level of coordination among agencies of the Federal government that are involved in high performance computing. The agencies involved have already begun cooperation to meet this challenge in order to mount a sufficiently comprehensive program in support of U.S. competitiveness.

The remainder of this section describes the goals for each of the four components of the Program, the actions that will be taken to achieve these goals, and the responsibilities of each of the participating federal agencies. Although the Program builds on the existing research base, its goals extrapolate significantly from those of the base and will require significant additional funding. This funding is presented in Table 1 at the end of the plan, with the funding elements keyed to each component and action of the Program.

High Performance Computing Systems

Recommendation: The U.S. Government should establish a long range strategy for Federal support for basic research on high performance computer technology and the appropriate transfer of research and technology to U.S. industry. [*HPC Strategy*, 1987]

Goals

High performance computing systems consist of processors, memory, mass storage, input/output, and associated system software. The systems are characterized in this report by overall sustained performance on large problems. They will be designed so that memory capacities, storage sizes and input/output rates scale to provide sustained performance in proportion to processing.

The goal is to support the development of high performance computing systems which will be capable of sustaining trillions of operations per second on significant problems. The program builds upon present government supported efforts which have established U.S. leadership in developing large scale computer systems and the underlying component technologies. However, achieving and effectively exploiting this thousandfold improvement in performance will require developing a new technology base through a program of research in computer architectures, microelectronics and packaging, and associated systems software .

A primary objective of the plan is to assist the continued viability of domestic sources of high performance computers and their critical components that meet the requirements of U.S. industry and Federal programs, both civil and defense. The plan will focus Federally funded research and promote transfer of results between Federally-funded research programs and U.S. industry. This requires close collaboration of researchers in the nation's universities and government laboratories with industrial scientists and engineers. Government funding will also assist risk reduction in critical areas and will complement private capital in the computer market.

To date Federal investment in high performance computing systems has taken two forms: (1) purchase of early market and production model systems and (2) research and development which has the led to commercial high performance systems. In both cases Federal funding has reduced the R&D risk of the high performance computing systems for U.S. manufacturers. Because many foreign computer manufacturers are willing to accept greater R&D risk due to their financial environment, the Federal strategy has been important in maintaining U.S. technical leadership in high performance computing systems.

Early purchase has served several critical functions. It has often provided essential financial assistance and early technical feedback to manufacturers during production of

their first model of a new high performance computing system. The substantial computational resources provided by these early purchases have maintained the rapid technological advance of the U.S. in both civil and defense sectors, thereby supporting the nation's economic and military security. Early users have devised efficient ways of exploiting the capabilities of new systems, providing a performance characterization of the design. They have also furnished significant new concepts to be incorporated in the designs of succeeding generations. This information has often led to improvements leading to more viable commercial products.

For example, NASA and DOE encouraged industry to develop a more advanced supercomputer to meet their research needs, which resulted in the Cray 2 supercomputer. This powerful new machine with greatly expanded memory might not have achieved market acceptance without the identified high performance computing requirement and subsequent acquisition by these Federal agencies. DOE and NASA acquired the first Cray 2 systems, and their early experience showed the broader market for computational research the importance of memories of hundreds of million words and provided valuable feedback that led to engineering changes for better performance.

Federal research and development investment has facilitated advanced research partnerships between industrial firms and university researchers. Industry provides practical knowledge and advanced manufacturing technology to produce high performance computing systems, while universities have developed new concepts and experimental systems. The results of these partnerships have often been significant in supporting the U.S. economy and national security.

The DARPA Strategic Computing program, DOE, and NASA have funded industry/university partnerships which have established U.S. leadership in scalable, highly parallel, high performance computing systems. Unlike the present generation of supercomputers, the resulting systems employ hundreds to thousands of processors. These architectures are generally scalable to higher levels of parallelism and, in the future, can exploit higher performance components and packaging with corresponding increases in sustained performance. This program has produced very promising results: the first generation of scalable parallel systems are now commercially available and have demonstrated high performance in both numeric and non-numeric applications. The second generation of this class of high performance computers is now emerging and scientific, engineering, defense, and business users are preparing for their arrival. Additional results include enhanced performance for workstations, personal computers, mass storage, graphics and input/output systems.

This Component of the HPC Program will build on recent experience in coordinated funding by different agencies of high performance computing systems research and development. For example, the Strategic Computing program at DARPA invested in R&D for an advanced parallel computer which was subsequently commercialized by

Thinking Machines Corporation. DARPA then collaborated with DOE and NASA to facilitate early use of this system in their research laboratories. The NSF recently funded a Science and Technology Center at Rice University, California Institute of Technology, Argonne National Laboratory, and Los Alamos National Laboratory which will consider more effective applications of this and other parallel architectures.

Action Plan

The focused high performance computing systems projects in this plan will be undertaken in cooperation with the software development projects. The systems also must be coordinated with advances in networking to ensure that their potential performance is available to remote users via the National Research and Education Network. The advanced research tasks provide excellent training grounds for the next generation of computer and computational scientists and engineers. Collaboration among these components is essential to the success of the Program.

Research for future generations of computing: Research in computer science, scalable parallel computer architectures, high density packaging technology, VLSI technology and optoelectronics will be increased. New packaging and component technologies will be developed together with associated design, analysis, simulation, and testing tools to enable their use in implementing larger scale computer architectures. This includes creating and extending models of computation, together with sufficient efforts in adaptation of fundamental algorithms, operating systems, and programming languages. These systems-specific activities complement the more generic and applications-focused software described in Advanced Software Technology and Algorithms where the emphasis is to develop the full potential of the new architectures.

System design tools: Support for rapid design, prototyping, and integration is essential to reach the capabilities needed for the Program. Progress in research, development and manufacturing of high performance computing systems is presently limited by lack of adequate automated design and analysis tools. A new generation of design tools and techniques will be developed for integrated, computer-assisted design and manufacturing of high performance computing systems from functional specifications through full systems. The tools will be developed so as to provide rapid prototyping in support of research, interfaced with the latest advances in automated manufacturing so as to boost U.S. industrial capabilities in addition to increasing research and development productivity. These facilities will make use of the latest high-density packaging technology which will be required to create systems at the targeted level of performance.

Transfer of technology to stimulate advanced prototypes: Cooperative university/industry high-risk research and development projects will provide rapid technology transfer from research results to working prototypes. Revolutionary concepts

are emerging from the frontiers of research in computer science and engineering, innovative computer architectures, mass storage systems, input/output systems, high density packaging, Very Large Scale Integration (VLSI) and optoelectronics. Government funds will be invested where opportunities exist for leverage to accelerate the transfer of the Federally-funded computing technology to American industry and vice versa. Advanced application-specific integrated circuits (ASICs) will be utilized where appropriate in these general purpose high performance computing systems. Joint projects in high risk areas will be pursued on a cost sharing basis with industry in close collaboration with government laboratories and academia. The focus of these projects will be to accelerate transition of high risk, revolutionary concepts from research laboratories into the commercial market while encouraging a domestic means of production for all critical components.

By the mid 1990s, it is expected that commercial advanced prototypes will be capable of sustaining two or three orders of magnitude better performance than today's systems for complex science, engineering, and defense applications, and for other problems of national importance. System software, including operating systems, programming languages, and software analysis tools, will be developed to determine the computational potential of the commercial systems. Performance analysis and measurement tools will be improved to enable the design and configuration of heterogeneous systems.

Evaluation of Early Systems: Evaluation of early production models of new high performance computing systems will be undertaken using representative problems. These systems will be acquired at the smallest scale that can evaluate their potential performance. The resulting evaluations will form a basis for decisions to develop the associated generic software and specific large scale applications in the Advanced Software Technology and Algorithms component. Needs of the Grand Challenges will be considered fully, and some early production models of high performance computers may be utilized in one or more of the Grand Challenges, at the sites performing this research.

This component of the HPC Program does not include acquisition of full scale systems. Some of these will be acquired under the High Performance Computing Research Centers in the Advanced Software Technology and Algorithms component; others will be purchased by Federal agencies to fulfill their missions. This investment will sustain the U.S. competitive edge and must be protected by ensuring appropriate export controls.

Rationale

Improvements in materials and component technology are advancing computer capability rapidly. Memory and logic circuits are continuing to improve in speed and density, but as fundamental physical limits are approached, advances are being sought through improved computer architectures, custom components, and software and

algorithms. Application-specific integrated circuits, such as for real-time signal processing, are being incorporated into special purpose computing systems. Computer architectures have begun to evolve into large scale parallel systems. Scalable architectures provide a uniform approach that enables a wide range of capacity, from workstations to very high performance computers.

At current performance levels our ability to model many important science, engineering, and economic problems is still limited. Computational models which have been developed for these problems require for realistic solutions speeds of trillions of operations per second and corresponding improvement in memory size, mass storage, and input/output systems. *Achievement of this performance level in the next five years is feasible, based on extrapolations of processor capability, demonstrated architectures, number of processors, and improved software performance.*

Responsibilities

NSF, NASA, DOE and DOD share responsibility for long-range research on the foundations of high performance systems. Within the DOD this responsibility will rest with DARPA, the Army Research Office (ARO), the Office of Naval Research, (ONR) and the Air Force Office of Scientific Research (AFOSR). These agencies have all been involved in this area and have considerable knowledge of the status and opportunities.

DARPA will carry the prime responsibility for high-risk research and development leading to commercialization of highly parallel high performance computing systems and will work with ARO, ONR, and AFOSR to achieve this end. DARPA will also have the lead responsibility for supporting research facilities for rapid design, prototyping, and integration of these systems, using advanced components and packaging. DARPA's unique style of managing high risk, large scale projects is particularly effective for transfering technology in joint university and industrial efforts.

The DOE, NASA and DARPA will continue to acquire first production models of high performance computing systems. The diversity of interests represented by these agencies has been important to the broad range of systems developed by industry in the U.S.. This healthy arrangement will continue.

NIST will expand its program for development of measurement techniques and performance modeling for high performance computer systems, and will support transfer of this technology to industry.

Advanced Software Technology and Algorithms

Recommendation: The U.S. should take the lead in encouraging joint research with government, industry, and university participation to improve basic tools, languages, algorithms, and associated theory for the scientific Grand Challenges with widespread applicability. [*HPC Strategy*, 1987]

Goals

Sustained improvements in computing hardware performance and sophistication have resulted in a shift from hardware and architecture to software and algorithms as the primary determiners of the power, flexibility, and reliability of major computing systems. Today the ability to exploit computing technology to address scientific and technological problems of competitive and national importance is determined primarily by software capability.

Breakthroughs in software technology enable computer solutions to problems whose scale, complexity or evolving nature previously inhibited any organized approach. Breakthroughs in algorithm design improve problem solving performance by orders of magnitude, making tractable computational solutions in problem areas where previously no solutions of any sort, or only traditional analytical or experimental methods, were available.

The goal for the Software Technology and Algorithms component of the High Performance Computing Program is to develop a base of software technology and algorithms that (1) will enable solution of Grand Challenge application problems in science and engineering, and that (2) will have broad national impact on software productivity and on systems capability and reliability.

The approach taken is to develop the advanced algorithms and software technology required to address applications problems on the scale of Grand Challenges, while ensuring that the generic technology developed can be applied to a broad range of computational problems. This investment may lead to the development of commercial products, but only after the new concepts have been illustrated and their feasibility demonstrated. Therefore, specific investments will be made to reduce the risks associated with the transition and adoption of these advanced technologies.

The U.S. lead in many areas of science and technology will be closely linked to advances made on important fundamental problems identified as Grand Challenges. Grand Challenges come from many fields from basic science to applied technology. Their solutions will have significant, national-level impact across diverse fields of interest to many Federal agencies. Appendix A describes several Grand Challenges and the agencies concerned with their solutions.

Improvements in algorithm design and implementation are as important to total "user realized" system performance as are performance improvements in the computer

systems in which these algorithms will be executed. High performance computing offers scientists and engineers the opportunity to simulate conditions that are difficult or impossible to create and measure. This new paradigm of computational science and engineering offers an important complement to traditional theoretical and experimental approaches, and it is already having major impact in many areas. New approaches combining numeric and symbolic methods are emerging. Development of new instruments and data generation methods in fields as diverse as genetics, seismology, and materials is accelerating demand for computational power. As problems grow to the size and complexity of Grand Challenges, and as computer architectures grow more complex in order to provide increased computing power, the software and algorithms challenge becomes significantly greater.

Effective exploitation of the performance potential of the emerging parallel systems poses a special challenge both to software technology and to algorithm design. The required software technology has many dimensions, ranging from systems software, advanced compilers, and languages, to programming environments for developing and adapting software, to large scale distributed data repositories. Also included are techniques for analyzing and constructing software with high reliability and numerical accuracy, design of high performance algorithms for solving generic problems on specific architectures, and development of algorithmic and software architectural approaches specific to solving the Grand Challenges.

Research in fundamental parallel algorithms is needed to provide a sufficient base of algorithms for high performance architectures. The characteristics of the generic algorithms are often strongly dependent on the computational model embodied in a particular machine architecture. Various models of parallelism yield different algorithms, as do heterogeneous systems configurations involving hybrid computational models.

Networking technology will also have significant influence on the design of algorithms for distributed systems. Fundamental algorithms must be specialized and combined to provide application-specific algorithms appropriate for the Grand Challenges. Algorithm design, development, optimization, and validation requires substantial resources and collaboration. Experimental facilities are a critical tool for developing and demonstrating applications and systems software, computer architectures, and networks.

Action Plan

Support for Grand Challenges: A principal focus of activity will be providing advanced software technology support to research groups collaborating to address the Grand Challenges. The purpose of this is not to provide sustaining support for this research, but rather to provide a means to reduce the risks assumed by Grand Challenge researchers when adopting innovative high performance computing technologies.

3. Program Plan:
Advanced Software Technology and Algorithms

Collaborative groups will include scientists and engineers concerned with Grand Challenge areas, software and systems engineers, and algorithm designers. These groups will be supported by shared computational and experimental facilities, including professional software engineering support teams, linked together by the National Research and Education Network. Groups may also create a central administrative base, which can be located anywhere on the network. Experimental facilities, often called testbeds, are included in the network in order to provide real-time access to data streams and support for rapid validation of computational models.

Technical contributions arising from this investment will include development of application-specific codes for innovative high performance computing systems, design and analysis of algorithms for Grand Challenge problems, and architecture and performance assessment as it relates to specific applications.

Agencies will select Grand Challenge applications to be included in this Program on the basis of the national importance of the specific area and the extent of cost-sharing from sources directly concerned with the specific scientific and engineering applications. An additional consideration will be the leveraging potential in other areas, in particular the commercial domain. Investment related to high performance computing will complement the traditional sources of support for Grand Challenge research by enabling exploratory use of advanced computational techniques.

Software components and tools: The Grand Challenge applications groups will have common needs in many areas of software technology including programming environments for code development and adaptation, advanced compiler technology. Also needed will be tools for optimization and parallelization, data management and interoperability, analysis and performance measurement, user interaction and visualization, debugging, and instrumentation. Advances in these generic software technology areas will have broad national impact, beyond the immediate scope of the Grand Challenge applications.

In order to provide these tools in a manner that is responsive to the needs of the applications researchers, collaborative groups will be formed that cut across the Grand Challenge areas in order to coordinate and share supporting software technology. This will enable multiple applications groups to sustain more easily a fast pace of innovation in the underlying software technology. These groups will include industrial, academic, and government researchers. Innovative approaches will be used to provide incentive for industry to participate and share costs.

A major focus of systems design and engineering will be developing advanced software applications that exploit the high capacity of the National Research and Education Network to provide new capabilities to researchers. An important example is a distributed operating system that permits high capacity interactions among programs at multiple network sites. This capability will enable a researcher to develop applications that may involve several high performance computers located at diverse sites to work

together effectively. Other applications include distributed shared data and program libraries, research report dissemination systems, and advanced user interaction and visualization systems. For example, security support and data interoperability are required to enable distributed databases that exploit the National Research and Education Network.

Computational techniques: Developing software tools and components is basic to fundamental research in computational technology. It is this research that yields the fundamental algorithms, models of computation, new approaches to program analysis, and language approaches that provide fundamental generational advances.

Research in computational techniques includes the areas of parallel algorithms, numerical and mathematical analysis, parallel languages, and program refinement techniques. Also included are models of computation, formal methods for high assurance, theoretical and empirical techniques for algorithm analysis, and related areas.

Results in design and theory of algorithms are as important to breaking down computational scaling barriers as are performance improvements in computing hardware. Algorithm breakthroughs continue to be made on even fundamental problems such as linear algebra that are often assumed to be well understood. Breakthroughs can yield thousandfold speedup factors above and beyond hardware advances, as illustrated in Figure 3.

Parallel computing is the principal source of opportunity to improve computational performance. There are many differences among the models of computation embodied in parallel computers, and all of these differ from the purely sequential model that dominated the first half-century of computing. Algorithm theory has already yielded scalable parallel solutions to many computational problems that were assumed by most practitioners to be inherently sequential. In order to realize the potential for performance and scaling implicit in the parallel computer technology, research in the design of algorithms will be supported.

The evolution of parallel computing technology has also stimulated renewed activity in the area of high level programming languages. Languages that have inherently sequential semantics force programmers to make unnecessary and undesirable computational commitments that must, in any case, be undone by optimizing compilers. Efforts will be funded to develop higher level languages that will enable computational scientists to consider separately the abstract computational problem being solved and the specific implementation approach. This will also enable use of emerging programming tools that integrate program transformation and optimization with analysis to yield implementations with higher assurance and predictable numerical characteristics.

High performance computing research centers: The HPC Program will support deployment of innovative high performance computing architectures to computational

scientists and engineers working on Grand Challenge applications, and to other computer scientists and engineers. Centers will be established to accelerate transition to new generations of high performance computing technology by enabling researchers to explore applications of this new technology.

Information gained from evaluating prototypes of new architectures as part of the High Performance Computing Systems Component will aid the choice of architectures to support the Grand Challenges. As the risks associated with application of innovative systems diminishes in individual Grand Challenges, the costs associated with the facilities will be transferred to the interested sponsors of the applications research. In this component we include computing hardware, network access to the operating systems of scientific instrumentation, and operational support for the Grand Challenge cooperative groups.

Facilities will also be provided to researchers in computing technology in order to support a more rapid transition to the new technology base. Researchers in areas such as algorithms, software environments, and operating systems require experimental access to new generation hardware. For example, there are a number of theoretical models for parallel computation in general use among algorithm designers, but only through empirical work can these models by adjusted to reflect more faithfully the models embodied in the parallel systems. Crucial systems parameters, for example, the relation of processing time to communications time and memory speed, interact with algorithm design parameters in ways that can best be explored empirically.

It is expected that many of the facilities allocated as part of Advanced Software Technology and Algorithms component will be used to facilitate transition of Grand Challenge applications to the new high performance computing systems. The remaining portion will be provided to computing technology researchers in order to support the development of generic algorithms and software technology. These facilities are in addition to those which will be provided as part of the High Performance Computing Systems component of the Program as a means to accelerate the transition from prototypes into products.

Responsibilities

DOE, NASA, NSF, NOAA and **DARPA** share responsibility for clearly defining the computational requirements of the Grand Challenges. They will select Grand Challenge applications areas in which collaborative groups are to be formed, and are responsible for providing advanced software technology support to the research groups collaborating to address the Grand Challenges applications in their domains.

NASA will carry lead responsibility for organizing and chairing the Federal Advanced Software, Technology and Algorithms Coordinating Committee. DOE, NSF, DARPA and other DOD research activities will be among the agencies participating in this committee.

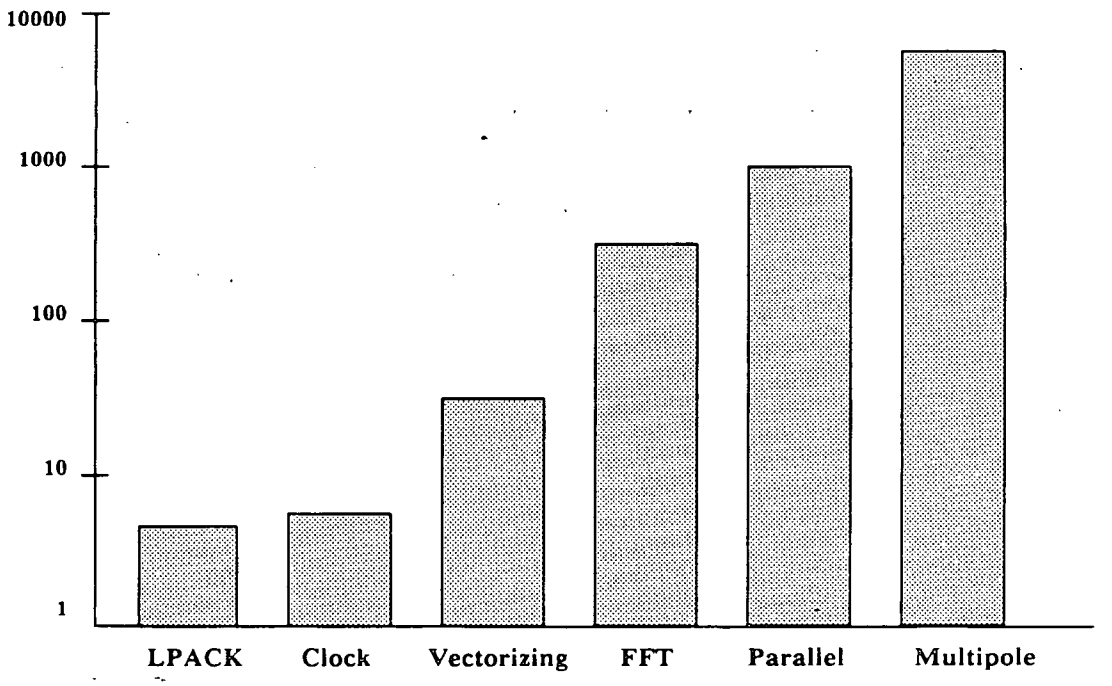
3. Program Plan:
Advanced Software Technology and Algorithms

NASA, DARPA, NSF and **DOE** will support development of software tools and standard components for use across the spectrum of the Grand Challenges. DOE and NASA share responsibility for exploiting the nearer term potential for commercialization of these software developments.

DOE, NASA, NSF and **DARPA** will incorporate early production models of the high performance computing systems into high performance computing laboratories. These high performance computing laboratories will include the advanced software tools and components, innovative computational techniques and the application-specific algorithms and experimental code for the Grand Challenges. These facilities will support the required integrated research, and will be available to users through the National Research and Education Network.

NSF, DOE, NOAA and **NASA** will build on their existing supercomputer centers which will provide the facilities for several high performance computing research centers, accessible to the national research community.

NOAA will be responsible for organizing the coordination of R&D in data management, and will play a lead role in supporting basic research in tools and techniques required for management and analysis of large-scale scientific data bases and distributed data handling.



Factors in Computational Speedup

Examples using vectors of length n=4096 (2^{12})

Method	Type	Order	Speedup
LAPACK: BLAS1->BLAS3 {Dongarra, et al.}	Algorithm & Software	constant	4
Hardware Clock Speed {1976 to 1989}	Microelectronics	12ns->2 ns	6
Vectorizing Compilers {CFT, VAST, KAP}	Algorithm & Software	32*scalar	32
Fast Fourier Transform {Cooley and Tukey}	Algorithm	$O(n^2)$->$O[n*\ln(n)]$	340
Parallel Processors {Gustafson, Montry and Benner}	Algorithm & Architecture	Linear in nr. of processors	1000
Fast Multipole Method {Greengard & Rokhlin}	Algorithm	$O(n^2)$->$O(n)$	4000

Fig. 3 - Speedup Due to Advances in Algorithms

The National Research and Education Network

Recommendation: U.S. Government, industry and universities should coordinate research and development for a research network to provide a distributed computing capability that links the Government, industry, and higher education communities. [*HPC Strategy*, 1987]

Introduction

The United States must develop a National Research and Education Network (NREN) to support communication between persons and organizations involved in open research and scholarly pursuits in the United States. This need has become increasingly obvious to the research community, especially among those who have experienced the benefits of electronic mail and database access, exchange of files between computers, and remote access to specialized and high-performance computing systems. As networking technology grows in power, network-based collaboration continues to allow substantive improvements in research effectiveness. These themes are well expressed in the recent National Research Council report *Toward a National Research Network* (1988). In developing the plan for this component the growing importance of the interrelationships between the network, the research components of the Program, and the U.S. academic community became increasingly clear. *Education* has been included in the name of the network in explicit recognition of this importance.

Today, all major organizations and government agencies use computer networking to some extent, and those with the most progressive and demanding missions have organized major transcontinental networks. A number of these networks are interconnected, notably those of the National Science Foundation (NSFNET), the Department of Defense (ARPANET and MILNET), the Department of Energy (ESNET), and the National Aeronautics and Space Administration (NASA Science Internet). These and many other commercial and regional networks collectively form the Internet, which currently supports a large portion of the U.S. science and engineering research community.

Today's Internet is far from uniform in the type and quality of service provided, and it does not yet reach the entire research community. Even so, expanding the Internet and enhancing its performance as far as technology allows will fall far short of what can and should be accomplished. The goal of this component is to create a new NREN which operates at rates of gigabits per second nationwide. This tremendous challenge is within the grasp of the United States in the next ten years. A network with this level of performance will provide another major improvement in the effectiveness of the national research community and their resulting ability to contribute to U.S. competitiveness.

3. Program Plan:
The National Research and Education Network

Availability of the NREN will provide an environment which enhances collaboration both for software technology development and for basic research and scholarship nationally. In return the development of the NREN will benefit from advances in software technology, particularly in the area of network services.

The eventual impact of the NREN on national competitiveness may well extend beyond such gains in research productivity. The NREN should be the prototype of a new national information infrastructure which could be available to every home, office and factory. Wherever information is used, from manufacturing to high-definition home video entertainment, and most particularly in education, the country will benefit enormously from deployment of this technology.

Stages of the NREN: The stages of NREN development as articulated in the *HPC Strategy* are:

The *first stage* involves an upgrade of the existing Internet to 1.5 megabit per second trunks. (This process is underway.)

The *second stage* will deliver upgraded network services to 200 to 300 research installations, using a shared backbone network with 45 megabit per second capacity.

The *third stage* will deliver one to three gigabit per second networking service to selected research facilities, and 45 megabit per second networking to approximately 1000 sites nationwide.

The stages of the NREN are illustrated in Figure 4.

Government/Industry/University roles: The Federal government plays a dual role in the development of computer networking. Federal funding has supported networking research and technology development in academic, industrial, and (to a lesser degree) government laboratories. The government also supports operational networks and network services. These are expected eventually to create a commercially viable market whose needs can be supplied by the private sector. In this latter role, the government has supplied networks as value-added services on communication circuits leased from the common carriers, and has subsidized their use by segments of the scholarly and research communities.

Universities play a major research role in advanced networking technology. Whereas most of the improvements in *communications* technology have come from industry, many of the most important *networking* technologies have been developed by universities. Educational institutions are also the primary users of networking nationwide, both for access to high performance computing and for collaboration among themselves and with government and industry.

To date, the role of industry has mostly been to provide communications links and produce equipment for networking. This situation is changing and in fact must be

radically altered in order to develop the high speed networks of the future. At data rates of gigabits per second the switching elements of the network need to be integrated with the communications links within the facilities of the communications industry. Applications of networks within and between industrial groups should also increase to support a more competitive U.S. industrial posture.

It is anticipated that the government will continue to fund networking research in partnership with academia and industry, and will continue to support parts of the national research networking infrastructure which do not yet have a sizable market. This will be necessary both to build the market for private offerings as well as other commercial goals. It also will be necessary for those government agencies sponsoring development of advanced networking to coordinate the work of multiple government laboratories, industrial, and university groups.

Action Plan

The Federal Research Internet Coordinating Committee (FRICC), a collaboration of the NSF, DARPA, DOE, NASA, and the Department of Health and Human Services (HHS), has begun transforming the present day Internet toward the goal of an NREN. This is being accomplished through sharing communications circuits, network access points, and even entire networks, leading to streamlined operations and reduced costs. The FRICC has established coordinating members in other agencies and national networking organizations and has developed a program plan for implementing the NREN. While these activities have provided a healthy start for the NREN, an additional effort will be necessary to achieve the ultimate goals of the High Performance Computing Program. FRICC, while not formally a part of the FCCSET structure of OSTP, works closely with the Committee on Computer Research and Applications and conducts its activities consistent with the policy guidance of the HPC Program.

Interagency effort to produce an interim NREN. Coordinating an interagency project as large as the NREN will not be easy. It is clear that a unified focus for management is necessary. It is equally clear that the project will not be fully supported by the diverse agencies involved unless they have a decisive role in shaping the project, and are kept in constant, close communication so that the resulting network fills their needs.

In *Stage 1* the agencies will continue to upgrade their networks to 1.5 megabit per second (T1) trunks. This effort is already well underway. In addition, DARPA project known as the *Research Internet Gateway (RIG)* is acquiring a prototype platform for development of "policy-based routing" mechanisms which will allow interconnection of these trunks. Also the FRICC has plans to develop enhanced capabilities such as directory services in support of network users.

As the Internet expands, issues of *network security* have become a source of increasing concern. Recent incidents have demonstrated the vulnerability of computers attached to

national networks. A significant effort in implementing the NREN will be development and implementation of mechanisms to enhance the security of the connected computing systems, and mechanisms to protect the networks themselves. These mechanisms will rely on policy-based routing capabilities, and also on recent advances in public-key cryptography.

In *Stage 2* the agencies are planning to acquire a common set of 45 megabit per second transcontinental trunks, the *Research Interagency Backbone (RIB)*. The ability to share backbone trunks, resulting in lower costs and improved service for all agencies, will be enabled by gateways with policy-based routing capabilities. When the RIB is fully operational, it will be interconnected with the NSFNET backbone; the result will be the interim NREN. Another equally important result will be the stage 2 technologies, which will provide a base from which commercial providers can offer compatible networking services nationally.

Research and development for billions of bits per second (gigabits) net. The ultimate structure of the *Stage 3* network will not become clear until this research effort is complete. However, it is clear that fiber-optic trunks now being installed by communications carriers will become increasingly important, new switching systems and network protocols must be developed, new high-speed interconnections to workstations and supercomputers will be needed, and some form of interconnection with the Stage 2 network will be needed. An additional goal of stage 3 is to support such advanced capabilities as remote interactive graphics, nationwide data files, and network-based high definition displays for education. Managing the dynamics of these activities will be a major challenge, but the payoff for success in terms of national capabilities will be enormous in terms of research productivity and, subsequently, in the form of technologies and services available from commercial sources.

Deployment of gigabits NREN. Stage 3 culminates in an operational national network with gigabits trunks. Deployment is not expected to begin until the middle to late 1990's.

Structured transition to commercial service: Mid-level networks organized on a regional basis or by other limited constituencies have sprung up indigenously (for example BITNET and several state-funded networks). Other mid-level networks have been formed with seed funding from NSF, NASA, and DARPA. These have become, in varying degrees, part of the existing Internet. They provide an important vehicle for the economic participation of state and local governments and of industry by providing access to the national network and by giving these other sectors a stake in its operation, thus reducing the funding burden on the Federal government. Moreover, each of these networks is typically a private and autonomous (although possibly subsidized) business entity; thus elements of the emerging national network have already become part of the private sector. Continuation of this trend will result in

opportunities for many companies to become involved in leading-edge data communications.

By the end of Stage 2, it is expected that every university and major laboratory will be connected to the NREN through a mid-level network. Present regional offerings vary widely in reliability and scope. To provide homogeneous and universal networking service, interaction of the Federal government with mid-level networks must increase. It is also to be expected that competition and other market forces will come into play between these networks.

Each of the services developed for the NREN must become available commercially at the earliest practical time. The intention is that networking infrastructure should be a commercial offering nationwide. The government and its contractors would then purchase network connections from companies which would provide service to subscribers in general.

Eventually, computer networking should be as pervasively available as telephone service is today. The corresponding ease of inter-computer communication will then provide the benefits associated with the NREN to the entire nation, improving the productivity of all information-handling activities. To achieve this end, the deployment of the Stage 3 NREN will include a specific, structured process resulting in transition of the network from a government operation to a commercial service.

Agency Responsibilities

NSF will be the lead agency for deploying the operating NREN within the HPC Program. NSF has assumed responsibility for supporting a backbone for the NREN, and will coordinate collaboration among Federal agencies in this area. The NSF role of support and coordination will expand as the NREN grows; NSF will upgrade and extend the operational network, providing advanced network services, and collaboration technology. NSF will also support and participate in the interagency networking testbed.

DARPA will be the lead agency for the Program's advanced networking technology research and development. DARPA's research leading to the advanced networking technology for gigabit speeds (Stage 3) will take place within its Command, Control, and Communications programs as the primary contribution of the Department of Defense to the NREN. DARPA will also create a testbed, jointly funded with other participating agencies, for advanced network technology and inter-agency collaboration.

DOE will provide networking support for the energy research community and participate in the interagency networking testbed.

NASA will provide networking support for the aerospace research community, participate in the interagency testbed, and support research on aerospace applications and technology with a focus on telescience research and development.

3. Program Plan:
The National Research and Education Network

DARPA, NSF, DOE and NASA will continue their active roles in governance of the Internet, and will expand these roles by providing representatives on the council which sets policy for the NREN.

NOAA will provide networking in support of the climate and global change research community and will participate in the interagency testbed.

NIST will participate by establishing networking standards, with particular emphasis on protocols and security standards. NIST will continue its traditional role of coordinating developing technologies such as Broadband ISDN with service providers, computer manufacturers, telecommunication manufacturers, system integrators and end users through the standards process.

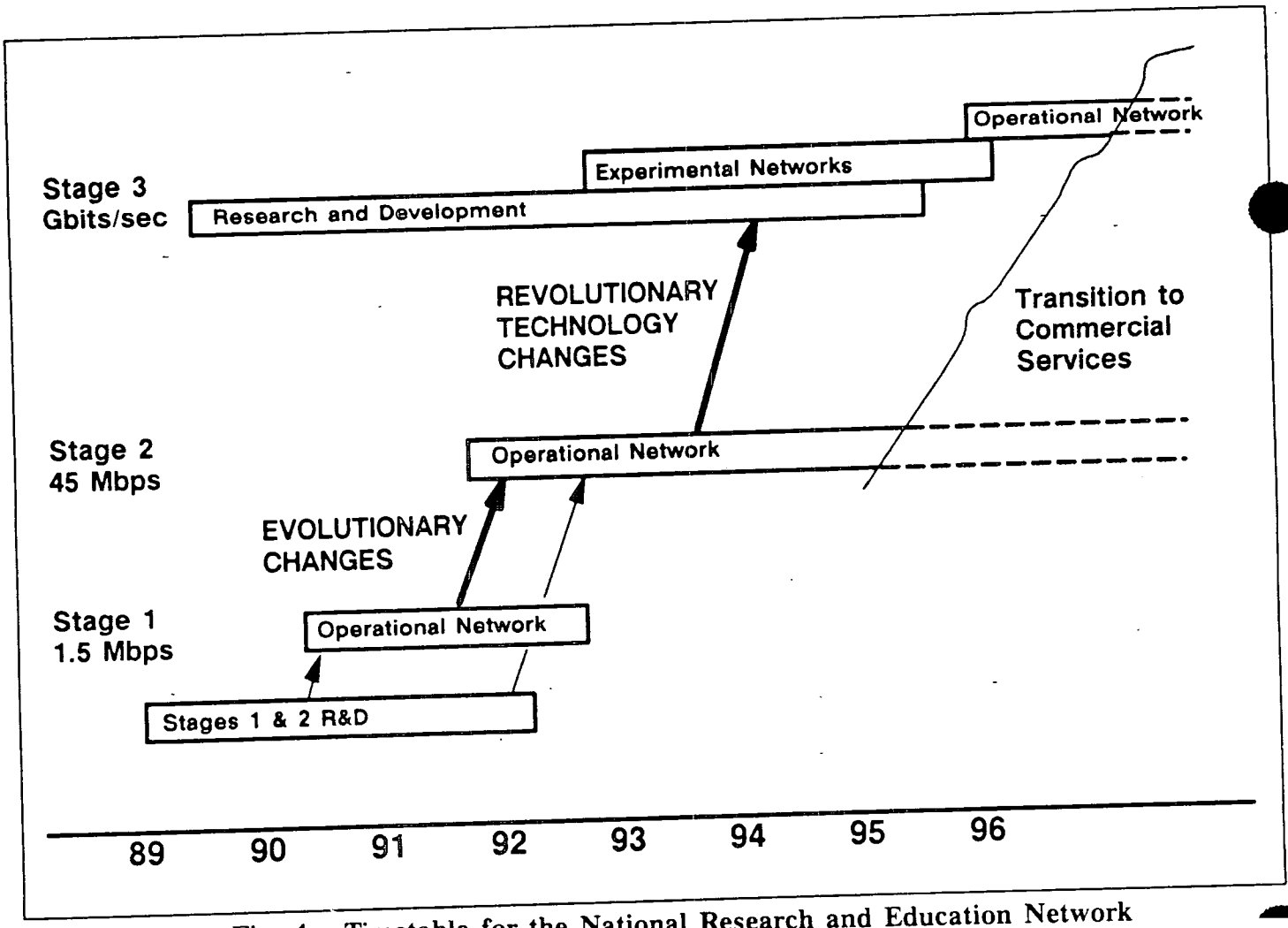


Fig. 4 - Timetable for the National Research and Education Network

Basic Research and Human Resources

Recommendation: Long term support for basic research in computer science should be increased within available resources. Government, industry, and universities should work together to improve the training and utilization of personnel to expand the base of research and development in computational science and technology. [*HPC Strategy*, 1987]

Goals

This component of the High Performance Computing Program addresses longer term national needs for high performance computing. The rapid growth of computing technology and computational science and engineering has created extraordinary demands for more rapid innovation, significantly increased manpower, and accelerated transfer of technology. The basic research community plays a major role in addressing these needs.

We must sustain a rapid pace of innovation in computer science and in computational science by investing in long term basic research.

Proprietary control is difficult to retain in an industry that is characterized by periodic major structural innovations, such as the shift now in progress from central timeshared computing to distributed networked workstations. Because of this, industry has little incentive to invest in long term approaches unless costs and risks are shared. For this reason, most of the major breakthroughs in computing have been the result of basic research activity. Examples include timesharing, local and national networks, VLSI design technology, personal computing, parallel computing, artificial intelligence, and many others. Each of these breakthroughs has had an enormous impact in the marketplace.

Increased numbers of qualified people are needed both in computational science and engineering, and in computer science and engineering. Universities are beginning to create new academic programs in areas of computational science and engineering that develop skills in both computer technology and in specific areas of science and engineering. The rapid evolution of this technology requires practitioners with a broad range of generic skills.

There is a need to reduce risk to industry in adapting and applying new technology. Technology is transferred rapidly from research into practice when the research community is an active participant in the process and when there is consensus in the research community on basic technical issues. The diversity of parallel computing models and algorithmic approaches now emerging provides unusual opportunity for application. The rapid pace of technology development in computing and computational science demands active participation of basic researchers in accomplishing transfer of the emerging technologies.

3. Program Plan:
Basic Research and Human Resources

Support for basic research must be supported in several respects. The High Performance Computing Program addresses this need through direct basic research support, improved infrastructure to increase research productivity, and facilitation of collaboration. The goals of this component of the Program are:

- **Basic research.** Ensure an adequate level of basic research activity to produce the next generation of innovative results in computing technology.
- **Human resources.** Support basic research, education, and training in order to meet the needs of research, personnel, and transition support in both computing technology and in computational science and engineering.
- **Support for collaboration.** Promote collaborations involving the basic research community, industry, and government to allow attacks on larger scale problems and accelerate dissemination of results.
- **Infrastructure.** Supprt the effectiveness of the research community by providing facilities and research infrastructure, including experimental high performance computers, networks, associated systems software, and applications software components.

The Basic Research Enterprise. Basic research programs already underway and supported by current Federal funding provide a base from which many computing applications goals can be achieved, but this base is already under great pressure even without the demands of important new thrusts[1]. The pace of expansion of computing technology and its applications greatly exceeds the rate of expansion of basic research, with consequent strain on the basic research community[2].

Effectively integrating new high performance computing technology into the US technological and scientific mainstream will require sustained research effort across the spectrum of computing technology. Some examples are microsystems component technology and packaging, computer architecture, fundamental algorithms and complexity, software engineering languages and tools, networking and distributed computing, artificial intelligence, numerical algorithms, and applications-specific algorithms.

Human Resources. Several studies in the past 10 years have documented the human resources challenges to the continued development and exploitation of computer technology[3]. A particular focus of these studies has been the severe undersupply of computer scientists and computer engineers at advanced degree levels. Computational scientists and engineers are in even shorter supply.

1. *The National Challenge in Computer Science and Technology*, Computer Science and Technology Board of the National Research Council, National Academy Press, Washington, 1988.
2. Gries, D., et al. *Imbalance Between Growth and Funding in Academic Computer Science: Two Trends Colliding*, Communications of the Association for Computing Machinery, 29(9), Sept 1986.

3. Program Plan:
Basic Research and Human Resources

Addressing the Grand Challenge applications requires large scale collaborative effort involving diverse groups of scientists, engineers, and mathematicians. The manpower shortage in computing technology and in computational science and engineering is hindering progress in these areas.

Collaboration. Interactions among multiple research teams and potential technology recipients contribute significantly to reducing the risks associated with transfer of major technologies into production. Collaboration can be facilitated in the basic research community by ensuring a high level of access to the network applications software based on the National Research and Education Network and by involving basic research groups in Grand Challenge applications.

The network support software will include capabilities for activities such as rapid distribution and sharing of research results, software distribution and configuration management support mechanisms, high-capacity interaction support for remote computers, access to instrumentation in remote experimental laboratories, rapid search and retrieval in distributed library databases, and so on.

Infrastructure. Scarcity of funds in computing technology research has hindered modernization of university computing research and education facilities. A rapid pace of technological innovation requires aggressive investment to ensure that universities remain at the forefront. Networked access to high performance computing with advanced software support is important for training computational scientists and engineers. The potential for using networks to disseminate results and conduct collaborative research at all educational levels is just beginning to be realized.

Basic infrastructure for computer research has been a concern for some years. Several agencies have programs that support the needs of the HPC Program. The Institutional Infrastructure Program at NSF has helped to equip approximately 25 computer science and computer engineering departments in the past eight years. DARPA for many years has been instrumental in building a core research base at major universities. The University Research Instrumentation program of the Department of Defense has provided important equipment and research support. The DOE research program has provided modern parallel computer facilities to several of its national laboratories and universities to promote basic research in high performance computing and to provide training facilities for graduate students and young faculty in all the disciplines involved in computational sciences. NASA maintains several research institutes and centers of excellence to interface with universities.

3. Gries, *op. cit.*; Feldman, Jerome A. and William R. Sutherland. *Rejuvenating Experimental Computer Science*. Comm. ACM 22, 9 (Sept. 1979), 497-502.; Kosaraju, S. Rao., et al. *Meeting the Basic-Research Needs of Computer Science*, Study report of the NSF CCR Advisory Committee, December, 1986; *Profiles -- Computer Science: Human Resources and Funding*, National Science Foundation report 88-324, February, 1989.

3. Program Plan:
Basic Research and Human Resources

Action Plan

Several specific approaches are taken to address the goals. These approaches will, in most cases, be implemented as possible expansions of existing research funding programs.

Expand basic research. Increase basic research activity in computing technology areas that influence high performance computing, including algorithms, software languages and tools, architectures, systems software, microsystems, networks, distributed computing, and symbolic processing.

Attain a level of 1000 computer science Ph.D.s per year by 1995. Strengthen the human resource infrastructure for basic research. Support university risk investment in computational science and engineering degree programs. This should be done by expanding the number of universities capable of providing high quality advanced education in computer and computational science and engineering.

Promote at least 10 computational science and engineering degree programs. Sponsor interdisciplinary programs in universities to accelerate the maturing of computational science and engineering subdisciplines.

Upgrade 10 university computer science departments toward the standards of current 10 best. Include facilities for research in high performance computing. Also, upgrade an additional 25 computer science departments to nationally competitive quality.

Provide National Research and Education Network access for every U.S. university and major laboratory. Every university and major laboratory will be connected to the National Research and Education Network through a mid-level network.

Improve facilities available to support basic research and advanced education. High performance computing facilities currently available to researchers are in such demand that there is only limited availability for educational usage in computational science and engineering degree programs. The effective introduction of computational science and engineering techniques into industry requires students to receive exposure to high-performance machines.

Improve ties between computing technology and other disciplines. Many breakthroughs in computational science and engineering applications result from interactions with computer scientists. Correspondingly, computer science, through exposure to the needs of computational science and engineering, is producing technology to address future needs. Funding will be directed to promote these interactions.

Provide access to professional engineering support. Professional engineering staff should be available to basic research groups for assistance in construction and maintenance of large scale prototype systems, including software and hardware. This

can be accomplished through industrial collaborations or through placement of professional staff in university laboratories.

Responsibilities

Federal agencies historically have supported activities which advance basic research by developing and improving infrastructure for the nation's knowledge and human resource base in computing. The HPC Program will exploit existing mechanisms to meet needs in these areas.

The **DOE** has established advanced computational science research facilities at several national laboratories and universities. Although the DOE labs already maintain a strong university cooperation program, more needs to be done to provide closer ties with the academic computational science community below the top echelon. For example, an expansion of the summer program for high school students using national facilities would be beneficial in providing interested and trained students to the universities.

The national laboratories are ideal training centers for graduate students in the sciences because of the wealth of experience in solving real world problems. This environment makes a valuable addition to the training of new scientists and should be made available to many more senior graduate students, post-doctoral fellows and young faculty than is currently possible.

The **NSF**'s primary mission is broad support of basic research and human resources in science and engineering. NSF recently reorganized to support a new research directorate, Computer and Information Science and Engineering (CISE), to focus resources on computing as a strategic research area. CISE supports research grants for academic institutional improvement as well as research. The research community supported by CISE will be primary participants in this Program.

Several NSF Centers (Science and Technology Centers, Engineering Research Centers, and Supercomputer Centers) focus on topics central to the HPC Program. These Centers illustrate the type of university-industry and interagency programs which can be employed directly as testbeds and sources of high performance computing technology.

The five NSF National Supercomputer Centers, for example, have provided advanced hardware and software to advance the utility of computational science across an entire spectrum of researchers. More than 11,000 scientists at some 300 institutions have used these facilities during the past few years. The research facilities and advanced experimental systems developed under the Program can be made available broadly to the entire U.S. research community through the National Research and Education Network.

NASA supports leading-edge applications of high performance computing technology. It also supports development of computational science programs in universities.

3. Program Plan:
Basic Research and Human Resources

NASA Institutes (ICASE, RIACS, ICOMP, CESDIS) and Centers of Excellence (CASIS, ICLASS) provide settings at NASA Centers or at Universities where computer scientists and computational scientists can work together using state of the art equipment on a permanent or temporary basis (summers, sabbaticals, etc.) These programs would be for undergraduates, graduate students, postdoctorals, university faculty, and researchers from industry and government.

NASA has supercomputer facilities at several of its field centers (Ames, Goddard, Langley, Lewis, and Marshall). NASA has also established the Numerical Aerodynamics Simulation (NAS) Facility, which is a national facility for aerospace applications which operates not only state-of-the-art supercomputers, but advanced parallel computers like the Connection Machine. NASA has also established a significant Artificial Intelligence Laboratory in the Information Sciences Division at ARC. These facilities are used for a wide range of mathematical, algorithm, systems software, and computer architectural research. These facilities are available to NASA centers, institutes, and grantees, and to the aerospace community. Under this program NASA facilities will be expanded to include scalable testbeds to support interdisciplinary research which combines mathematics, algorithms, systems software, and computer architecture.

NOAA has supercomputer facilities, and has also created generic, broad spectrum workstation design facilities to support the Program for Regional Observing and Forecasting Systems (PROFS). Under this Component, NOAA will expand the opportunities for collaborative research at its facilities for development of algorithms and techniques for large scientific data bases, use of artificial intelligence in data management, and development of climate prediction models.

DARPA provides high performance computing systems for research community use on two scales: small-scale for experimentation, software and algorithm development by computer and computational science research groups, and medium-scale for shared-use facilities intended for access by dozens of groups via the National Research and Education Network. DARPA funding also supports key university and industrial labs for research and advanced development in computer and network architecture, network protocols and management, microsystems design and prototyping, advanced components and packaging, software tools and parallel algorithms.

4. Organization

Leadership of the HPC Program is the responsibility of the Office of Science and Technology Policy. It will be coordinated through the FCCSET Committee on Computer Research and Applications, whose members include representatives of the key agencies. The Committee will work closely with the President's Science Advisor and the various government funding agencies to ensure the continuing success of the Program. The components of the program that implement the Program will be executed by the cognizant agencies. Duties and responsibilities of the Committee include:

- Interagency planning and coordination;
- Policy development and technology assessment;
- Liaison with the industrial and university sectors; and
- Annual reporting of progress to the Office of Scientific and Technology Policy.

A High Performance Computing Advisory Panel will be formed, consisting of eminent individuals from government, industry, and academia. Members of the Advisory Panel will be selected by and will report to the Director of OSTP. The Panel will provide the Director and the Committee with an independent assessment of:

- Progress of the Program in accomplishing its objectives;
- Continued relevance of the Program goals over time;
- Overall balance among the Program components; and
- Success in strengthening U.S. leadership in high performance computing, and integration of these technologies into the mainstream of U.S. science and industry.

A broadly representative industry body will assist in making long-range demand and robustness projections for: high capacity research networks; the spectrum of computer architectures; the adequacy of software development; and the level of the manpower pool. This body will help assure a smooth transition between successive generations of high performance computing systems.

The FCCSET Committee on Computer Research and Applications has established subcommittees that will be responsible for planning, organizing, monitoring and coordinating the components of this Program. This includes liaison with the industrial and academic sectors, and published annual reports.

5. Budget

Budgets for the Program are presented in Table 1. Each budget element corresponds to a key activity in one of the four components of the HPC Program. The activities are described for each component in Section 3 of this plan. The yearly additional funding requested for this Program corresponds to the estimate given in the *HPC Strategy*, with some adjustment to yearly funding levels as a result of more detailed planning and inflation. Significant portions of the Program's funding will be allocated in each of the three participating sectors: universities, industry, and government laboratories.

Currently, the four principal funding agencies are spending about $500 Million per year on research and development for high performance computing. It is important that this funding continue with coordination by the FCCSET Committee as discussed in this plan, because the ability of the Program to achieve its goals depends upon maintenance of the broad base of computational and computer science and engineering research presently funded by the Federal government.

Preliminary planning estimates suggest that the first year of the program would require an augmentation of $150 million, which would then grow to an incremental annual level of $600 million by the fifth year.

Special attention has been devoted to the subcomponents "Early Systems for Evaluation" and "High Performance Computing Research Centers". There is an explicit strategy for investment in emerging high performance computing systems (including associated software) in these activities, to ensure that adequate funding is available. It is intended that the Early Systems for Evaluation budget sustain acquisition of the smallest scale systems which will allow characterization of their potential performance. For systems which prove to have good performance potential, the High Performance Computing Research Centers budget will support scaling these systems up, to demonstrate that potential in the Grand Challenges or other advanced applications. This will reduce risk to both producers of the systems and researchers using them, to provide the necessary incentive for early deployment in the most advanced applications.

The Basic Research and Human Resources component also requires special discussion, because it is funded in two ways. First, ten percent of the Program funding is set aside for this component. Second, it is intended that an additional fifteen percent of the total Program funding in the other three components will consist of basic research, carried out largely in Universities, which will also support the Program goals in Basic Research and Human Resources. Integrating this research with the rest of the Program allows a smooth flow of research from basic ideas through to applications.

Summary of Additional Funds

(Millions of Dollars)

Reference Page	Component	Year 1	Year 2	Year 3	Year 4	Year 5
17	**High Performance Computing Systems**	**55**	**91**	**141**	**179**	**216**
19	Research for Future Generations	11	17	24	32	37
19	System Design Tools	10	18	21	25	25
19	Advanced Prototype Development	22	36	65	86	116
20	Evaluation of Early Systems	12	20	31	36	38
23	**Advanced Software Technology and Algorithms**	**51**	**90**	**137**	**172**	**212**
24	Support for Grand Challenges	9	19	34	43	48
25	Software Components and Tools	15	30	41	60	78
26	Computational Techniques	6	10	18	19	31
26	High Performance Computing Research Centers	21	31	44	50	55
31	**National Research and Education Network**	**30**	**50**	**95**	**105**	**110**
33	Interagency Interim NREN	14	23	55	50	50
34	Gigabits Research and Development	16	27	40	55	60
34	Deployment of Gigabits NREN		(Funding begins after Year 5)			
34	Structured Transition to Commercial Service		(Funding begins after Year 5)			
37	**Basic Research and Human Resources**	**15**	**25**	**38**	**46**	**59**
		NOTE:	15% of the other three Components is also commited to this general area			
	TOTAL High Performance Computing Program	**151**	**256**	**411**	**502**	**597**

Table 1 - Budget Summary by Program Component

ACKNOWLEDGMENTS

Office of Science and Technology Policy guidance was provided by William R. Graham, Tom Rona, Robert Post, and Paul Huray. The Plan Drafting Committee was chaired by David B. Nelson from the Department of Energy; and included J. Mark Pullen, William L. Scherlis, and Stephen L. Squires from the Defense Advanced Research Projects Agency; Donald Austin, Daniel Hitchcock, Thomas Kitchens, and Norman H. Kreisman from the Department of Energy; Paul H. Smith from National Aeronautics and Space Administration; and Melvyn Ciment, Peter Freeman, and Stephen Wolff from the National Science Foundation. Special thanks are due to Sandy Merola and Dennis Hall from the Lawrence Berkeley Laboratory, for editing and production assistance.

APPENDIX A: SUMMARY OF GRAND CHALLENGES FOR WHICH SOLUTION IS LIKELY TO BE POSSIBLE USING SYSTEMS DEVELOPED UNDER THIS INITIATIVE

PREDICTION OF WEATHER, CLIMATE, AND GLOBAL CHANGE. The aim is to understand the coupled atmosphere, ocean, biosphere system in enough detail to be able to make long range predictions about its behavior. Applications include understanding CO2 dynamics in the atmosphere, ozone depletion, climatological perturbations due to man made releases of chemicals or energy into one of the component systems, and detailed predictions of conditions in support of military missions.
Agencies: DOE, DOD, NASA, NSF, NOAA

CHALLENGES IN MATERIALS SCIENCES. High performance computing has provided invaluable assistance in improving our understanding of the atomic nature of materials. These have an enormous impact on our national economy. A selected list of such materials includes: semiconductors, such as silicon and gallium arsenide and superconductors such as the high Tc copper oxide ceramics that have been shown recently to conduct electricity at about 100 degrees Kelvin.
Agencies: DOD, DOE, NSF, NASA

SEMICONDUCTOR DESIGN. As intrinsically faster materials, such as gallium arsenide are used, a fundamental understanding is required of how they operate and how to change their characteristics. Essential understanding of overlay formation, trapped structural defects, and the effect of lattice mismatch on properties are needed. Currently, it is possible to simulate electronic properties for simple regular systems, however, materials with defects and mixed atomic constituents are beyond present capabilities.
Agencies: DOD, DOE, NSF

SUPERCONDUCTIVITY. The discovery of high temperature superconductivity in 1986 has provided the potential of spectacular energy-efficient power transmission technologies, ultra sensitive instrumentation, and devices using phenomena unique to superconductivity. The materials supporting high temperature superconductivity are difficult to form, stabilize, and use, and the basic properties of the superconductor must be elucidated through a vigorous fundamental research program.
Agencies: DOE, NSF, DOD

STRUCTURAL BIOLOGY. The function of biologically important molecules can be simulated by computationally intensive Monte Carlo methods in combination with NMR of crystallographic data. Molecular dynamics methods are required for the time dependent behavior of such macromolecules. The determination, visualization, and analysis of these 3D structures is essential to the understanding of the mechanisms of enzymic catalysis, recognition of nucleic acids by proteins, antibody/antigen binding, and many other dynamic events central to cell biology.
Agencies: DOE, HHS, NSF

DESIGN OF DRUGS. Predictions of the folded conformation of proteins and of RNA molecules, by computer simulation is rapidly becoming accepted as a useful, and sometimes primary tool in understanding the properties required in drug design.
Agencies: DOE, HHS, NSF

HUMAN GENOME. Comparison of normal and pathological molecular sequences is our current most revealing computational method for understanding genomes, and the molecular basis for disease. To benefit from the entire sequence of a single human will require capabilities for more than three billion subgenomic units, as contrasted with the ten to two hundred thousand units of typical viruses.
Agencies: DOE, HHS, NSF

QUANTUM CHROMODYNAMICS. In high energy theoretical physics, computer simulations of QCD are yielding first principle calculations of the properties of strongly interacting elementary particles. New phenomena have been predicted including; the existence of a new phase of matter, and the quark-gluon plasma. Properties under the conditions of the first microsecond of the big bang, and in the cores of the largest stars have been calculated by simulation methods. Beyond the range of present experimental capabilities, computer simulations of grand unified "theories of everything" have been devised using QCD (Lattice Gauge Theory).
Agencies: DOE, NSF

ASTRONOMY. Data volumes generated by Very Large Array (VLA) or Very Long Baseline Array (VLBA) radio telescopes currently overwhelms the available computational resources. Greater computational power will significantly enhance their usefulness in

exploring important problems in radio astronomy, resulting in better return on a major national investment.
Agencies: NASA, NSF

CHALLENGES IN TRANSPORTATION. In the nearer term, substantial contributions to vehicle performance can be made using more approximate physical modeling and reducing the amount of interdisciplinary coupling. Examples include, modeling of fluid dynamical behavior for three dimensional flow-fields about complete aircraft geometries, flow inside of engine turbomachinery, duct flow, and flow about ship hulls.
Agencies: NASA,DOD,DOE,NSF,DOT

VEHICLE SIGNATURE. Reduction of vehicle signature (acoustic and electromagnetic, and thermal characteristics) is critical for low detection military vehicles.
Agencies: NASA, DOD

TURBULENCE. Turbulence in fluid flows impacts the stability and control, thermal characteristics, and fuel performance of virtually all aerospace vehicles. Understanding the fundamental physics of turbulence is requisite to reliably modeling flow turbulence for the analysis of realistic vehicle configuration.
Agencies: NASA, DOD, DOE, NSF, NOAA

VEHICLE DYNAMICS. Analysis of the aeroelastic behavior of vehicles, as well as the stability and ride analysis of vehicles are critical assessments of land and air vehicle performance and life-cycle.
Agencies: NASA, DOD, DOT

NUCLEAR FUSION. Development of controlled nuclear fusion requires understanding the behavior of fully ionized gasses at very high temperatures under the influence of strong magnetic fields in complex three dimensional geometries.
Agencies: DOE, NASA, DOD

EFFICIENCY OF COMBUSTION SYSTEMS. To attain significant improvements in combustion efficiencies requires understanding the interplay between the flows of the various substances involved and the quantum chemistry which causes those substances to react. In some complicated cases the quantum chemistry required to understand the reactions is beyond the reach of current supercomputers.
Agencies: DOE NASA, DOD

ENHANCED OIL AND GAS RECOVERY. This challenge has two parts: to locate as much of the estimated 300 billion barrels of oil reserves in the US and then to devise economic ways of extracting as much of this as possible. Thus improved seismic analysis techniques as well as improved understanding of fluid flow through geological structures is required.
Agencies: DOE

COMPUTATIONAL OCEAN SCIENCES. The objective is to develop a global ocean prediction model incorporating temperature, chemical composition, circulation, and coupling to the atmosphere and other oceanographic features. This will couple to models of the atmosphere in the effort on global weather as well as having specific implications for physical oceanography.
Agencies: DOD, NASA, NSF, NOAA

SPEECH. Speech research is aimed at providing a communications interface with computers based on spoken language. Automatic speech understanding by computer is a large modeling and search problem in which billions of computations are required to evaluate the many possibilities of what a person might have said within a particular context.
Agencies: NASA, DOD, NSF

VISION. The challenge is to develop human-level visual capabilities for computers and robots. Machine vision requires image signal processing, texture and color modeling, geometric processing and reasoning, as well as object modeling. A competent vision system will likely involve the integration of all of these processes with close coupling
Agencies: NSF, DARPA, NASA

UNDERSEA SURVEILLANCE FOR ASW. The Navy faces a severe problem in maintaining a viable anti-submarine warfare (ASW) capability in the face of quantum improvements in Soviet submarine technology, which are projected to be so substantial that evolutionary improvements in detection systems will not restore sufficient capability to counter their advantages. An attractive solution to this problem involves revolutionary improvements in long-range undersea surveillance which are possible using very high gain acoustic arrays and active acoustic sources for ASW surveillance. These methods will be computationally intensive; even taking advantage of inherent parallelism and judicious design of algorithms, computational demands for the projected post-2000 era submarine threat mandate achieving signal processing computation rates of in excess of a trillion operations per second.
Agencies: DOD

APPENDIX B: GLOSSARY

bits- binary digits (the smallest units of digital information); also an abbreviation for "bits per second"

broadband ISDN (BISDN)- broadband integrated services data network; an evolving standard commercial communications offering which will provide data rates of hundreds of megabits per second

byte- one character of computer storage

common carrier- a regulated commercial company which offers communication services in an open market

flops- abbreviation for floating-point operations per second, a unit which characterizes the performance of a computer for certain scientific and engineering calculations

giga- prefix meaning billion, e.g. "gigaops" means "billion operations per second" and "gigabits" means "billion bits per second"

links- long-distance communications circuits, also known as "trunks"

mega- prefix meaning million, e.g. "megaops" means "millions operations per second" and "megabytes" means "million characters of storage"

mid-level network- a computer network with scope which falls between a nationwide network and a local network, such as one of the state or regional networks

ops- abbreviation for "operations per second", a general measurement of computer performance

policy-based routing- a computer network function which treats data packets in different ways depending on some policy, for example certain packets may be given high priority, certain others may be rejected as not authorized to use some portion of the network

telescience- science practiced at a distance, using telecommunications

tera- prefix meaning trillion, e.g. "teraops" means "trillion operations per second"

testbed- an configuration intended to allow experimentation with systems in an application environment

trunks- long-distance communications circuits, also known as "links"

value-added services- services provided in addition to basic communication links (and at extra cost); for example, computer networking using communications provided by a common carrier

FCCSET COMMITTEE ON COMPUTER RESEARCH AND APPLICATIONS

Paul Huray (Chair)
Office of Science and Technology Policy

SUBCOMMITTEES

Science and Engineering Computing

David B. Nelson (Chair)
Department of Energy

James Burrows
National Institute of Standards and Technology

Melvyn Ciment
National Science Foundation

Harlow Freitag
Supercomputer Research Center

Clarence Geise
Strategic Defense Initiative Office

Randolph Graves
National Aeronautics and Space Administration

Thomas Kitchens
Department of Energy

Norman H. Kreisman
Department of Energy

Jacob V. Maizell, Jr.
National Institutes of Health

C. E. Oliver
Air Force Weapons Lab

John P. Riganati
Supercomputing Research Center

Paul B. Schneck
Supercomputing Research Center

K. Speierman
National Security Agency

Paul H. Smith
National Aeronautics and Space Administration

Stephen L. Squires
Defense Advanced Research Projects Agency

Computer Research and Development

Jacob T. Schwartz (Chair)
Defense Advanced Research Projects Agency

Donald Austin
Department of Energy

James Burrows
National Institute of Standards and Technology

Bernard Chern
National Science Foundation

Peter Freeman
National Science Foundation

Lee Holcomb
National Aeronautics and Space Administration

Charles Holland
Air Force Office of Scientific Research

Robert E. Kahn
Computer Science Technology Board

Daniel R. Masys
National Institutes of Health

Robert Polvado
Central Intelligence Agency

David Sadoff
Department of State

William L. Scherlis
Defense Advanced Research Projects Agency

K. Speierman
National Security Agency

Stephen L. Squires
Defense Advanced Research Projects Agency

Andre Vantilborg
Office of Naval Research

Computer Networking, Infrastructure and Digital Communications

William A. Wulf (Chair)
National Science Foundation

Peter Alterman
Health and Human Services

F. Ronald Bailey
National Aeronautics and Space Administration

Charles Brownstein
National Science Foundation

James Burrows
National Institute of Standards and Technology

Daniel Hitchcock
Department of Energy

Arnold Pratt
National Institutes of Health

J. Mark Pullen
Defense Advanced Research Projectes Agency

Rudi F. Saenger
Naval Research Laboratory

Anthony Villasenor
National Aeronautics and Space Administration

Stephen Wolff
National Science Foundation

HIGH PERFORMANCE COMPUTING AND COMMUNICATIONS:

Investment in American Competitiveness

Prepared for
U.S. Department of Energy,
and
Los Alamos National Laboratory

by
Gartner Group, Inc.
56 Top Gallant Road
Stamford, CT 06902

March 15, 1991

NOTICE:

The preparation of this report has been supported by funds provided by the U.S. Department of Energy through contracts with the Los Alamos National Laboratory. Hence, this report may be copied, excerpted, and/or quoted without restriction and without prior permission of the U.S. Department of Energy, the Los Alamos National Laboratory, or Gartner Group, Inc.

A limited number of copies of this report have been provided to the Los Alamos National Laboratory in accordance with contract provisions. Additional copies may be obtained from Gartner Group for a modest fee which covers the cost of reproduction, handling, and postage.

DISCLAIMER:

The judgements expressed in this report are solely those of Gartner Group, Inc. and do not necessarily reflect the views of the United States Government, the U.S. Department of Energy, the Los Alamos National Laboratory, or any of the persons who have contributed to this study or the organizations with which they are affiliated.

[This page has been left blank intentionally.]

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF EXHIBITS

TABLE OF EXHIBITS (cont'd)

TABLE OF EXHIBITS (cont'd)

TABLE OF EXHIBITS (cont'd)

[This page has been left blank intentionally.]

CHAPTER I - INTRODUCTION

I - INTRODUCTION

This report presents the results of a study performed by Gartner Group, Inc. for the U.S. Department of Energy (DOE), through a contract with the Los Alamos National Laboratory (LANL). The purpose of the study is to estimate the economic impact of the Federal High Performance Computing and Communications (HPCC) Program which was proposed by the Office of Science and Technology Policy (OSTP), Executive Office of the President, on September 8, 1989. That Program is an implementation of the **Research and Development Strategy for High Performance Computing**, which was transmitted to Congress by OSTP on November 20, 1988.

This introductory chapter describes the objectives of the study and presents the arrangement of subsequent chapters.

OBJECTIVE

The objective of this Gartner Group study is to provide an assessment of the likely economic impact and benefits of the Federal HPCC Program and the risks of non-support of this program. The goals of the HPCC Program are to:

- **Support computational advances through R&D effort;**
- **Reduce uncertainties to industry through increased cooperation and continued use of government as a market for High Performance Computing (HPC) prototypes;**
- **Support underlying research, network, and computational infrastructures; and**
- **Support the U.S. human resource base.**

OBJECTIVE

The HPCC Program will consist of four complementary, coordinated components in each of the key areas of High Performance Computing and Communications:

- **High Performance Computing Systems;**
- **Advanced Software Technology and Algorithms;**
- **National Research and Education Network; and**
- **Basic Research and Human Resources.**

This program will augment the existing Federal base funding for computer and information science and technology research and development, which now amounts to about $500 million per year, by $1.9 billion over a five-year period. In the first year of the program, about $150 million in additional funding will be provided, and this amount will increase each year, culminating with almost $600 million additional in the final year. About 35 percent of the total funding will be allocated to the High Performance Computing Systems component, slightly less to Advanced Software Technology and Algorithms, about 20 percent to the National Research and Education Network (NREN), and the remaining 10 percent to Basic Research and Human Resources.

For further details of the HPCC Program, see Appendix A.

APPROACH

This study was carried out in two phases:

- Phase I focused on the development of two alternative scenarios: one assuming that the HPCC Program is not supported, the other assuming that it is.

- Phase II focused on assessment of the near-term economic impact of the HPCC Program by comparing the effects of the two alternative scenarios upon users of High Performance Computing (HPC): in particular, the ability of various industries to realize product and market objectives which can only be attained through HPC.

The methodology employed in each of these phases is described in Appendix B.

ACKNOWLEDGEMENTS

A number of persons contributed to this study by providing advice, criticism, data, encouragement, expertise, leads, and other useful information. These persons are identified in Appendix C. It is virtually impossible to indicate precisely who provided exactly what, but we are very appreciative of the large amount of assistance we have received, as well as the cooperative spirit in which it was given.

However, Gartner Group is solely responsible for the contents of this report. The opinions expressed do not necessarily represent the views of the persons named or the organizations with which they are affiliated.

ARRANGEMENT OF THE REPORT

This report consists of six chapters, plus ten appendices, as follows:

I **Introduction** - (this chapter)

II **Executive Overview** - explains the present situation in High Performance Computing, presents the highlights of the two alternative scenarios to the year 2000, and summarizes the major differences between the two scenarios.

III **Background** - presents an introduction to High Performance Computing and describes HPC applications opportunities and obstacles.

IV **The HPC Arena** - reviews the developments in the supercomputer market over the past ten years and presents two scenarios for the coming decade:

- Scenario "A" assumes that business goes on as usual: the Federal HPCC Program is <u>not</u> supported.
- Scenario "B" assumes full support of the Federal HPCC program.

V **HPC Applications** - describes the projected impact of the Federal HPCC Program on American industrial competitiveness and technological leadership.

VI **Conclusions and Recommendations** - summarizes the findings of this study and presents recommendations for some additional steps to strengthen HPC in the United States.

Appendices A-J - present additional information in support of Chapters I-VI.

[This page has been left blank intentionally.]

CHAPTER II - EXECUTIVE SUMMARY

II - EXECUTIVE SUMMARY

This chapter summarizes the highlights of the report, and is arranged in five sections:

- **Overview**
- **Background (Chapter III)**
- **The HPC Arena (Chapter IV)**
- **HPC Applications (Chapter V)**
- **Conclusions and Recommendations (Chapter VI)**

OVERVIEW

The thrust of this report is that the United States should fund the Federal High Performance Computing and Communications (HPCC) Program as proposed by the Office of Science and Technology Policy (OSTP). Three basic lines of argument are used to support this position:

1. The U.S. leadership in High Performance Computing (HPC) is threatened by Japanese companies with deep resources and long-term outlooks. The HPCC Program is needed to maintain and enhance U.S. leadership in the supercomputer industry and in other related industries.

2. There are major opportunities for computational science applications that will significantly enhance U.S. industrial competitiveness, productivity, energy situation, quality of life, basic science, and national security. The HPCC Program will accelerate the realization of those opportunities.

3. Leadership in computer science and technology (especially supercomputers and related products) and leadership in computational science applications go hand-in-hand. It is impossible for a country to be a leader in the one field without also being a leader in the other.

BACKGROUND

Chapter III introduces the basic concepts and players in HPC. The principal points are as follows:

HPC Represents The "Leading Edge" In Information Technology - It is the part of the information industry where change is occurring most rapidly. Through the "Trickle-Down Effect," HPC affects the rest of the information industry, which in turn affects all (or nearly all) other industries. This is why the Federal HPCC Program is so important: perhaps nowhere else could the expenditure of a relatively small amount of government funds have so great an effect.

Japanese Supercomputer Companies Are Stronger Than U.S. Companies - Cray Research, the principal U.S. supercomputer vendor, has annual revenues of less than $1 billion, and the other U.S. vendors are under $200 million. By contrast, the three largest Japanese computer companies, Fujitsu, Hitachi and NEC, have annual revenues ranging from $17 billion to $45 billion, and they have demonstrated a willingness to subsidize supercomputing R&D. Furthermore, they provide one another the strong domestic rivalry that helps build global competitiveness. IBM, DEC and the other large U.S. computer companies have not been major contenders in the supercomputer business, although IBM is now showing signs of becoming one.

HPC Applications Are Critical - Not only is HPC at the frontiers of computing, but computational science, as delivered through HPC applications, is at the frontiers of science, engineering, and related endeavors. To fall behind in one arena is to invite falling behind in the other as well.

The "bottom line" of Chapter III is this:

- **U.S. government support is essential for computing technologies in their infancy, especially in light of increasing Asian and European R&D activities; and**
- **U.S. government assistance in technology diffusion is also imperative, given the close working relationships between government and industry in Japan and (to a lesser extent) in Europe.**

The barriers which must be overcome are:

- **Erroneous perceptions that HPC is inordinately expensive and difficult to use; and**
- **Lack of understanding, especially by management, of the potential benefits of using HPC (*i.e.*, computational science).**

Overcoming these barriers is what the Federal HPCC Program is all about.

THE HPC ARENA

Chapter IV begins by tracing the development of HPC from 1980 to 1990. The main messages are as follows:

HPC Growth Rates Have Been Substantial, But The HPC Market Is Still Relatively Small - The supercomputer industry has shown a compound annual growth rate of just under 30 percent in the last decade, about double that of the computer industry as a whole. While worldwide revenues have grown from $89 million in 1980 to over $1.1 billion in 1990, this still represents less than one percent of overall computer industry revenues. Consequently, supercomputer development is not served well by free market forces.

U.S. Leadership In HPC Is Declining - Japanese vendors have mounted a strong challenge to the U.S. in the race to build the fastest supercomputers . Cray Research, the leading supercomputer company, has seen its market share slip from 90 percent of the world's supercomputer systems in 1980 to slightly more than half in 1990, while Japan's market share has grown from zero in 1980 to 28 percent in 1990. The geographic distribution of installed supercomputer systems has changed as shown below.

Exhibit II-2: Distribution of Installed Supercomputers, 1980-1990

	1980	1990
United States	81%	50%
Europe	11%	19%
Japan	8%	28%
Other	–	3%

SUMMARY: THE HPC ARENA (Chapter IV)

Next, Chapter IV presents two alternative scenarios extending from the present to the year 2000:

- **Scenario A assumes "business as usual;"**
- **Scenario B assumes full funding and support of the Federal HPCC Program.**

The scenarios were developed from a variety of sources which, taken together, form a "jury of expert opinion:"

- Existing Gartner Group Scenarios;
- The Gartner Group Information Industry Model;
- Historical background;
- Assumptions (*i.e.*, wise allocation of funds under the Federal HPCC Program);
- Experience; and
- Peer Review.

The major common thread in Scenarios A and B is the gradual dominance of parallel supercomputers over vector supercomputers. In the latter 1990's, shipments of vector supercomputers will taper off, while parallel supercomputer shipments will accelerate. The major differences between Scenarios A and B are in the rate of improvement and penetration of parallel supercomputers and in the position of the U.S., relative to Europe and Japan, in the use of supercomputers. Exhibit II-3, following, presents these differences in terms of supercomputer technologies and markets in the year 2000.

Exhibit II-3: Comparison Of Scenarios A and B, Year 2000

	SCENARIO A (BUSINESS AS USUAL)			SCENARIO B (FUNDING OF HPC)		
	U.S. Vector Supercomputers	Japanese Vector Supercomputers	Parallel Supercomputers	U.S. Vector Supercomputers	Japanese Vector Supercomputers	Parallel Supercomputers
Installed Power	3.7 million megaflops (2%)	13.5 million megaflops (8%)	158.4 million megaflops (90%)	4.6 million megaflops (1%)	13.5 million megaflops (3%)	422.0 million megaflops (96%)
	(Total: 175.6 million megaflops)			(Total: 440.1 million megaflops)		
Average System Price	$25 million	$16 million	$35 million	$22 million	$16 million	$38 million
Average System Power	12 gigaflops	38 gigaflops	630 gigaflops	12 gigaflops	38 gigaflops	1,300 gigaflops
Installed Systems	640 (34%)	669 (36%)	552 (30%)	754 (35%)	669 (31%)	750 (34%)
	(Total: 1,861 systems)			(Total: 2,173 systems)		

From a usage standpoint, the differences between the two scenarios are shown below:

Exhibit II-4: Installed Supercomputers In The Year 2000

	SCENARIO A	SCENARIO B
United States	683	995
Europe	345	345
Japan	768	768
Other	65	65
TOTAL	1,861	2,173

This difference implies into a 28 percent increase in supercomputer revenues over the next decade for Scenario B as opposed to Scenario A. In "current" dollars, this increase amounts to:

$10.4 billion.

HPC APPLICATIONS

As described in Chapter V, the Federal HPCC Program will improve productivity in research and development (R&D) by enabling companies to:

- **Undertake development which would be impossible otherwise;**
- **Bring new products and services to market more quickly; and**
- **Develop better products and services.**

These capabilities are essential to maintaining or increasing industrial competitiveness.

This improved productivity in R&D equates to increased overall industrial productivity, at least in proportion to R&D expenditures. In the five key industrial sectors selected for analysis in this study, the HPCC Program is expected to increase productivity as shown in Exhibit II-5. Similar productivity improvements should also occur in other industrial sectors, depending upon their level of HPC usage.

Exhibit II-5: Annual Productivity Improvements, 1991-2000

Aerospace	0.50% to 1.98%
Chemicals	0.83% to 3.31%
Electronics	0.96% to 2.11%
Petroleum	0.88% to 4.79%
Pharmaceuticals	0.22% to 0.47%

Econometric modeling runs using the University of Maryland's Long-term Interindustry Forecasting Tool (LIFT) predict that, when extended to the economy as a whole, these productivity gains will have a salutary effect upon a number of economic indicators, as shown in the following exhibit.

Exhibit II-6: Economic Impact of the Federal HPCC Program

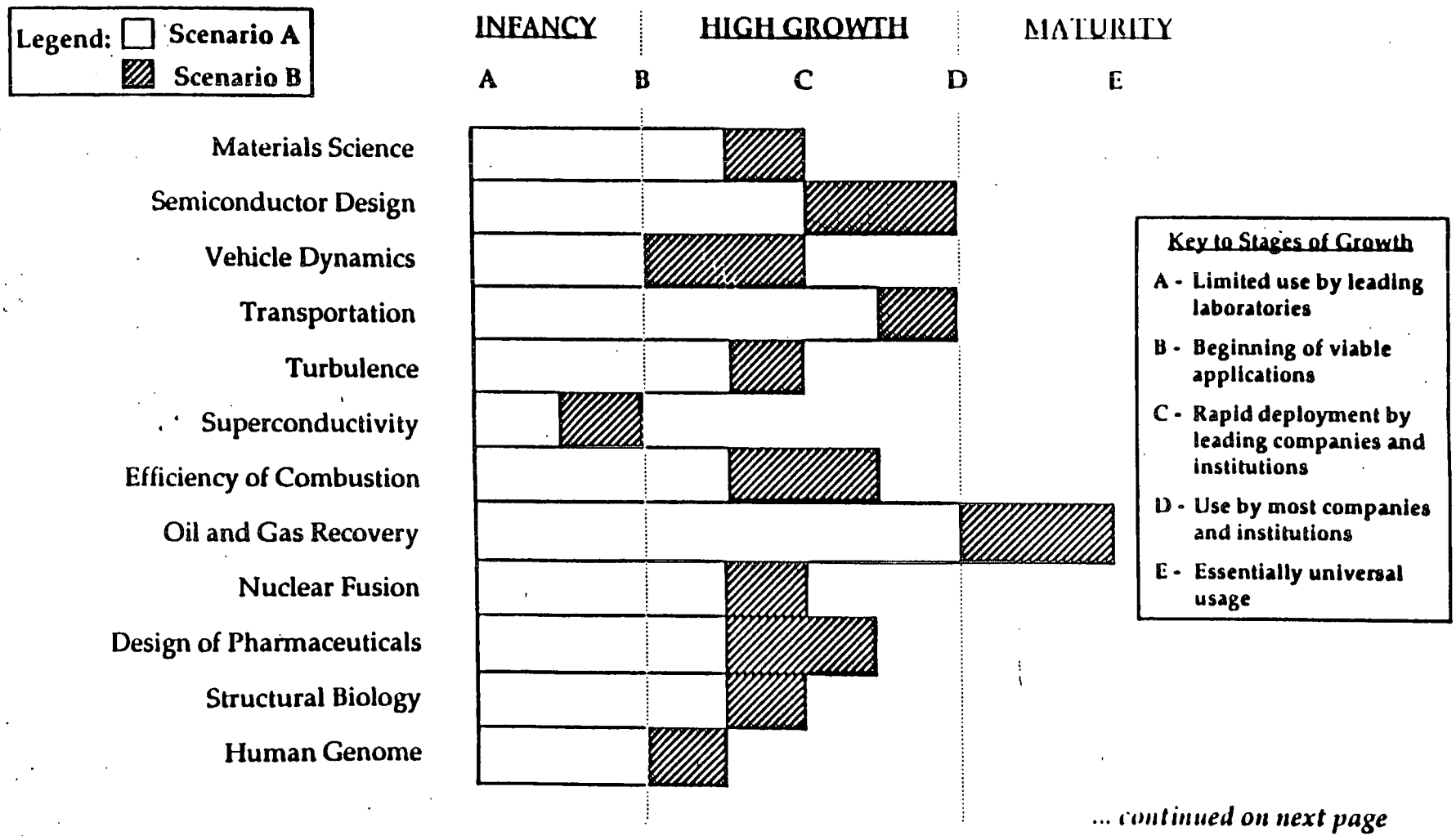
($ Billions, 1982 constant dollars)

	Year 2000 only	1991-2000 Cumulative
Gross National Product	up $28.9 to $83.9	up $172.5 to $502.6
Personal Consumption	up $16.2 to $44.4	up $101.8 to $280.6
Gross Private Domestic Investment	up $8.5 to $25.7	up $57.5 to $199.2
Gross Exports	up $3.4 to $12.5	up $8.4 to $30.6
Net Exports (less Imports)	up $4.2 to $13.8	up $3.2 to $22.8
Federal Deficit	down $13.0 to $30.8	down $74.7 to $190.3

In addition to these economic consequences, the Federal HPCC Program will benefit science and technology by improving the level and sophistication of HPC usage in a broad range of important applications. The extent of improvement can be determined by comparing the levels of application sophistication for Scenarios A and B in the following exhibit. The difference between the two scenarios is attributable to the HPCC Program.

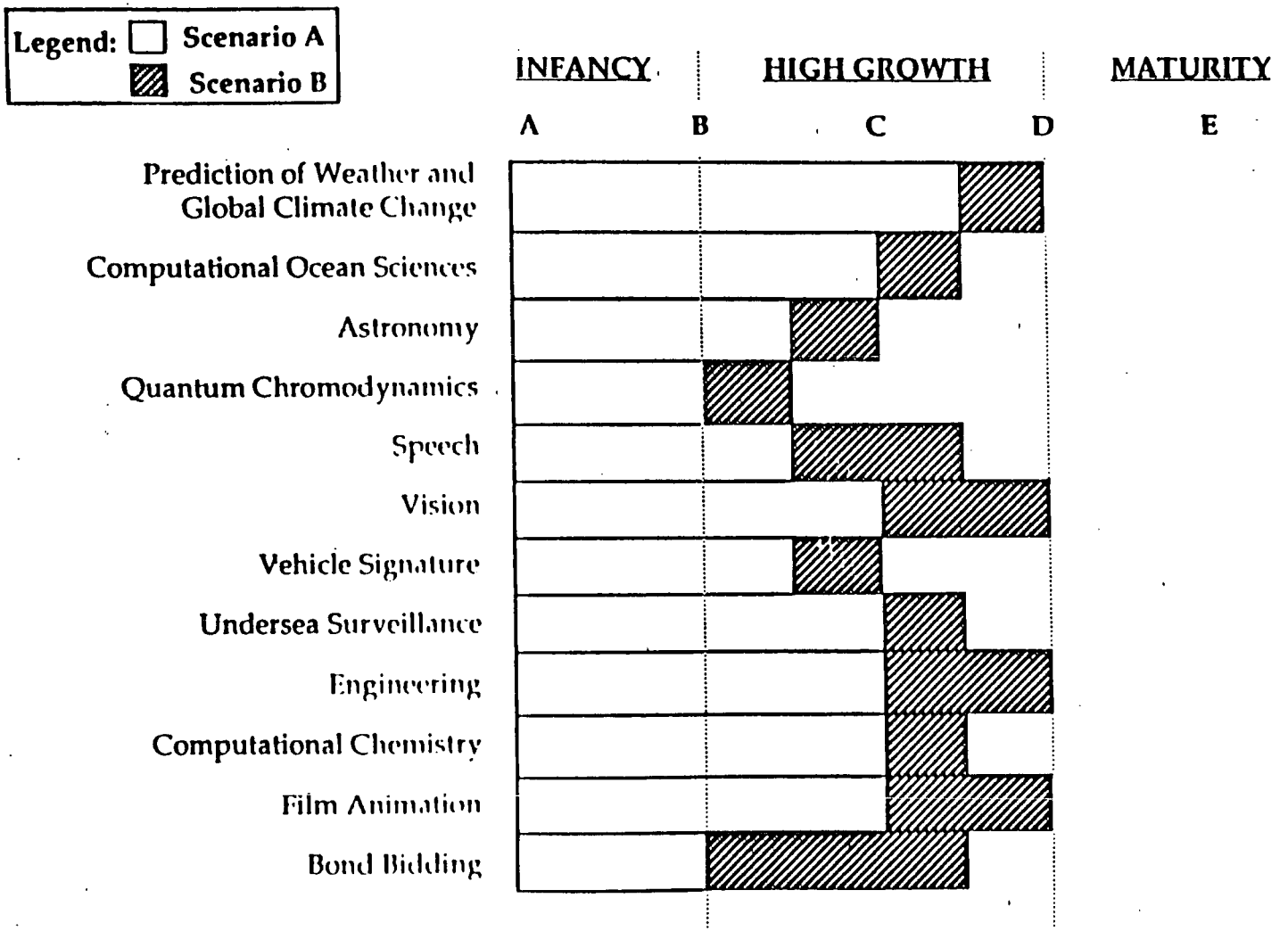
SUMMARY: HPC APPLICATIONS (Chapter V)

Exhibit II-7: Level of HPC Application Sophistication



... continued on next page

Exhibit II-7 (cont'd)



CONCLUSIONS AND RECOMMENDATIONS

In our judgement, the direct effects of the Federal HPCC Program will be threefold:

- **First, it will affect at least some of the directions of change in HPC;**
- **Second, it will affect the rate of change in HPC.**
- **Third (and probably most important), it will affect the rate of application of HPC throughout American industry, academia, and government.**

As shown in Chapters IV and V, the economic value of these effects, over just the next ten years, will be much, much greater than the amount of Federal funding projected for the Program.

Therefore, we recommend that the Program be inaugurated, with full funding, as soon as possible.

As a by-product of our study, we have also identified three areas in which the Federal HPCC Program could -- and, we believe, should -- be strengthened:

- **Technology Transfer;**
- **Alternative Education Programs; and**
- **Monitoring and Evaluation.**

These are explained in Chapter VI.

[This page has been left blank intentionally.]

CHAPTER III - BACKGROUND

III - BACKGROUND

This chapter provides an overview of High Performance Computing (HPC). It consists of six sections, as follows:

- **High Performance Computing** - defines what HPC is, at least for the purposes of this report.
- **HPC Technologies** - explains the basic technologies involved in HPC and why HPC plays a unique role in the information industry.
- **HPC Vendors** - describes in general terms the "supply side" of HPC.
- **HPC Usage** - summarizes past growth in supercomputer applications and reviews the major application opportunities in HPC.
- **Computational Science** - briefly explains what computational science is and its relationship to HPC and HPC applications.
- **Obstacles** - reviews the generic obstacles standing in the way of realizing HPC application opportunities.

HIGH PERFORMANCE COMPUTING

High Performance Computing (HPC) may be defined simply as use of the fastest, largest, and/or most advanced computers currently built. Traditionally, this has meant "supercomputers," a very special *genre* of computers that will continue to be the quintessence of HPC, but it has also come to include other classes of computers that embody at least some elements of the HPC "spirit": namely, "minisupercomputers," mainframes with special auxiliary processors, high performance workstations, computer systems with novel architectures, etc. It has also come to include the software and networks which make the high performance hardware accessible to users.

What these elements have in common is that they are extraordinary. They are at the cutting edge, rather than at the average level, of information technology in one or more aspects. Hence, HPC is important to the information industry in general because it is the frontier of advanced information technology. That alone makes HPC essential to any industrial (or post-industrial) nation, because of the importance of the information industry internationally. But HPC is even more important because of the extraordinary things it enables and facilitates in other industries and in other areas of science and technology. It permits the rapid development of new and improved products and services which would be impractical or impossible otherwise, and it opens the way to a whole new paradigm for scientific investigation.

HPC TECHNOLOGIES

The kinds of technologies which drive HPC are the same ones which drive information systems of all types:

- **Components,**
- **Architectures,** and
- **Software.**

Components are the building blocks of information systems: the semiconductor circuits used in processors and memory (see Appendix D for a brief discussion of HPC Terminology) and the other elements used in peripheral devices, workstations, networks, etc. **Architectures** determine the structure of information systems: which components are utilized, and how. They define what functions will be provided, which of these will be implemented in hardware, and which in software. **Software** specifies what an information system actually does in a given situation. It determines the "user interface": how the information system "looks and feels" to the user. It is the link between the system hardware and the application.

What distinguishes the components used in HPC systems from those used in other information systems is their capacity. In semiconductor circuits, for example, this can be measured in switching speed, gates (that is, circuit elements) per chip, heat dissipation per gate, etc. Circuit switching speed obviously affects how fast a computer can perform a given operation, so HPC system designers are constantly looking for faster circuitry from which to build new systems. The number of circuit elements per chip also affects system capacity, because it determines the complexity of functions that can be completed in a single machine cycle. Circuit density becomes a capacity limiter as supercomputer cycle times get faster and faster, because electronic signals can travel no faster than the speed of light. (Light travels about 11.8 inches in one nanosecond -- a nanosecond is one-billionth of a second -- and supercomputers with cycle times in the 2-3 nanosecond range are now becoming available.) However, denser circuits can cause heat dissipation problems, unless ways are found to attain rapid switching speeds at low power. (See Appendix D for a further discussion of components.)

HPC systems differ in architecture from "typical" information systems because of their emphasis upon performance: that is, solving a particular problem (or class of problems) as quickly as possible. Because of this, HPC systems may employ unusual architectural features which make them somewhat less "general-purpose" than most information systems. (There is a fundamental trade-off between being able to solve a wide class of problems reasonably well and being able to solve a particular problem very well.) Examples of this are "vector processing" and "large-scale parallelism." (Again, see Appendix D.) On the other hand, HPC researchers can sometimes attain greater performance in a particular situation by adopting and adapting architectural approaches that have been successful in other applications, so there is also an impetus to make HPC systems ever more "general-purpose."

Software can be evaluated in terms of system utility: that is, how easy (or difficult) is it for a user to get the system to solve a particular problem. The concept of utility is a difficult one, however, because it depends not only upon the problem (as does performance) but also upon the user, who is (or should be) capable of changing and learning. Although HPC systems have a reputation for being difficult to use -- perhaps deservedly so, because early supercomputers tended to sacrifice utility for performance -- that has changed considerably in recent years. Nowadays, HPC systems are at the leading edge in visualization -- showing the computational results as a still or moving picture, rather than page upon page of numbers -- which promises to greatly enhance the utility of all kinds of information systems in the coming years.

In all of these aspects -- components, architecture, and software -- HPC systems are somewhat like racing cars. They are rather specialized in purpose and form in order to attain levels of functioning which otherwise would not be possible. And like racing cars, it is difficult to say precisely what characteristic(s) distinguish HPC systems from the run-of-the-mill variety, but when a user moves from the latter to an HPC system, or a racing car, the difference is apparent. As with automobiles, high performance in computers is multi-dimensional: it can manifest itself as processor performance, memory capacity, input-output capability, or whatever, just as cars can be judged on the basis of acceleration, top speed, handling, braking, etc.

The "Trickle-Down" Effect

There is another important aspect in which HPC systems are like racing cars: what constitutes high performance in computers changes over time, what was "very high" performance just a few years (or even months) ago may be only mediocre today. The reason for this is that both HPC systems and racing cars serve as test environments for new concepts and technologies which, if proven successful, are employed in later generation "mainstream" situations. (Indeed, this is often a major reason why automobile manufacturers have auto racing programs, and some experts would claim it is a primary reason why Japanese companies have entered the supercomputing business.) Thus, computing concepts such as pipelining, multiprocessing, cache memory, and extended (semiconductor) storage, all of which were used only in supercomputer and "near-supercomputer" systems a mere decade ago, are now appearing in widely-used, commercially-oriented systems -- just as overhead camshaft engines, fuel injection, disk brakes, and independent suspension have moved from racing cars into family sedans.

Other instances of "trickle-down" from supercomputers are to be found in the mainframe systems recently announced by Fujitsu, Hitachi, and NEC. The central processors used in the Fujitsu M-1800 series, the Hitachi M-880 series, and the NEC ACOS System 3800 series are based upon the scalar portions of the Fujitsu VP-2000, the Hitachi S-820, and the NEC SX-3 supercomputers, respectively. And just as these supercomputers are threatening to overtake those made by industry-leader Cray Research as the "world's most powerful," their mainframe derivatives are challenging those made by IBM in terms of basic hardware performance (raw MIPS).

IBM, however, is expected to employ some "trickle-down" of its own within the next two years to maintain its position of overall leadership in the industry. For instance, it will introduce special hardware processors to boost the performance of its DB2 relational database subsystems, and it will make increasing use of parallelism to provide significant processing performance improvements in a wide range of applications. These approaches will be based upon architectural concepts used in the Vector Facility (which IBM introduced in late 1985 to enable its 3090 mainframe line to compete in supercomputer markets), the RS/6000 family of HPC workstations and departmental systems, and IBM's prototype HPC systems developed in its research laboratories.

This "trickle-down" phenomenon is one reason why High Performance Computing is so important, not just for a narrow, specialized group of users, but for the entire industry. The drive to achieve ever higher performance leads computer (and automotive) designers to take advanced concepts and techniques from the research laboratories and attempt to exploit them in an "almost real world" environment: HPC (or auto racing). The new approaches which survive in those crucibles are subsequently passed along to the "completely real world" in the ensuing years. (A conceptual framework which explains the role of HPC in advancing the general state-of-the-art in computing is presented in Appendix E.)

HPC VENDORS

Unlike automobile companies, most of the HPC systems makers in the U.S. are not "full-line manufacturers." Instead, they are "niche" players, making only supercomputers, minisupercomputers, or high performance workstations. IBM and DEC are the most notable exceptions, but DEC is not yet a serious player in the supercomputer market and IBM has moved into contention only recently. As a result, most American HPC vendors have a relatively small market base over which to amortize their considerable R&D expenses. (The cost of developing a new supercomputer system is currently about $100 million.)

The situation is quite different in Japan, however, as shown in Exhibit III-1. All of their supercomputer makers -- Fujitsu, Hitachi, and NEC -- make a full range of computer systems, extending from portable personal computers to general-purpose mainframes. Hence, the Japanese can (and do) regard supercomputers as "loss leaders": that is, product lines which are never intended to be profitable in and of themselves, but which contribute (through "trickle-down" and prestige) to the overall success of their maker companies. This, coupled with the other advantages that Japanese companies enjoy -- *e.g.*, lower profit margins, "patient" capital, lower interest rates -- puts most U.S. HPC vendors at a decided disadvantage relative to Japanese competition.

To make matters worse, all of the Japanese computer firms also make semiconductors for the worldwide merchant market, so they are able to underwrite the development of advanced circuits from a very broad base. Hence, it is no surprise that Japan now leads the world in advanced semiconductor technology: not only silicon-based chips, but also exotic new technologies such as gallium arsenide (GaAs), which may well become the preferred supercomputer component technology of the 1990s. Meanwhile, most U.S. semiconductor companies have been driven out of these expensive and very risky markets, leaving U.S. HPC vendors (other than IBM) to depend upon the Japanese -- notably their erstwhile supercomputer competitors -- for advanced components.

Exhibit III-1: Characteristics of Leading HPC Companies

Company	Total Revenues	HPC Revenues	R&D Expenditures	Comments
Cray Research	784.7	784.7	143.3	
Cray Computer	0.0	0.0	40.7	
CDC	2,934.5	26.0	289.2	Dropped out 4/89
IBM	62,710.0	631.9	6,827.0	
Alliant	67.9	67.9	11.3	
Convex	158.6	158.6	20.7	
Thinking Machines	45.0	45.0	15.0	
Fujitsu	16,627.9	127.4	1,708.3	Figures are for Japanese fiscal year ended 3/31/90. Conversion rate ¥150 = $1.00
Hitachi	45,139.2	58.4	2,682.5	
NEC	21,517.8	120.0	3,387.3	

All figures are for 1989, in U.S. $ millions.

Sources: **Datamation**, company reports, Gartner Group estimates

Nevertheless, the past decade has seen a number of new American entrants in the HPC business, especially in minisupercomputers and workstations. In many cases, these companies were leaders in bringing important new HPC technologies to market, and from time to time this has enabled them to wrest important sales from the grips of the industry leaders. However, the long-term prognosis for these companies is not particularly bright. Not only do they suffer from all of the disadvantages *vis-a-vis* the Japanese described above but they also face the inevitable problems stemming from an increasingly crowded market. As a result, some of these companies have disappeared almost as suddenly as they appeared, while others have died a slower death. The best hope for those that have survived into the 1990s may be to be bought out by a larger, more broadly based firm, possibly a Japanese one seeking quick access to new technology.

The prospective customers of these start-up companies are faced with a dilemma. On one hand, they are attracted by the new and attractively-priced technological capabilities which these companies invariably offer, but on the other hand, they are apprehensive about the vendor's chances of long-term survival. The sudden withdrawal of Control Data Corporation from the supercomputer business in 1989 exacerbated these fears and has even raised doubts about the ability of industry-leader Cray Research to endure in the long term. Ultimately, that will be determined largely by the market climate for HPC, both in the United States and internationally, and that in turn will be affected considerably by government actions such as the Federal HPCC Program or the lack thereof.

HPC USAGE

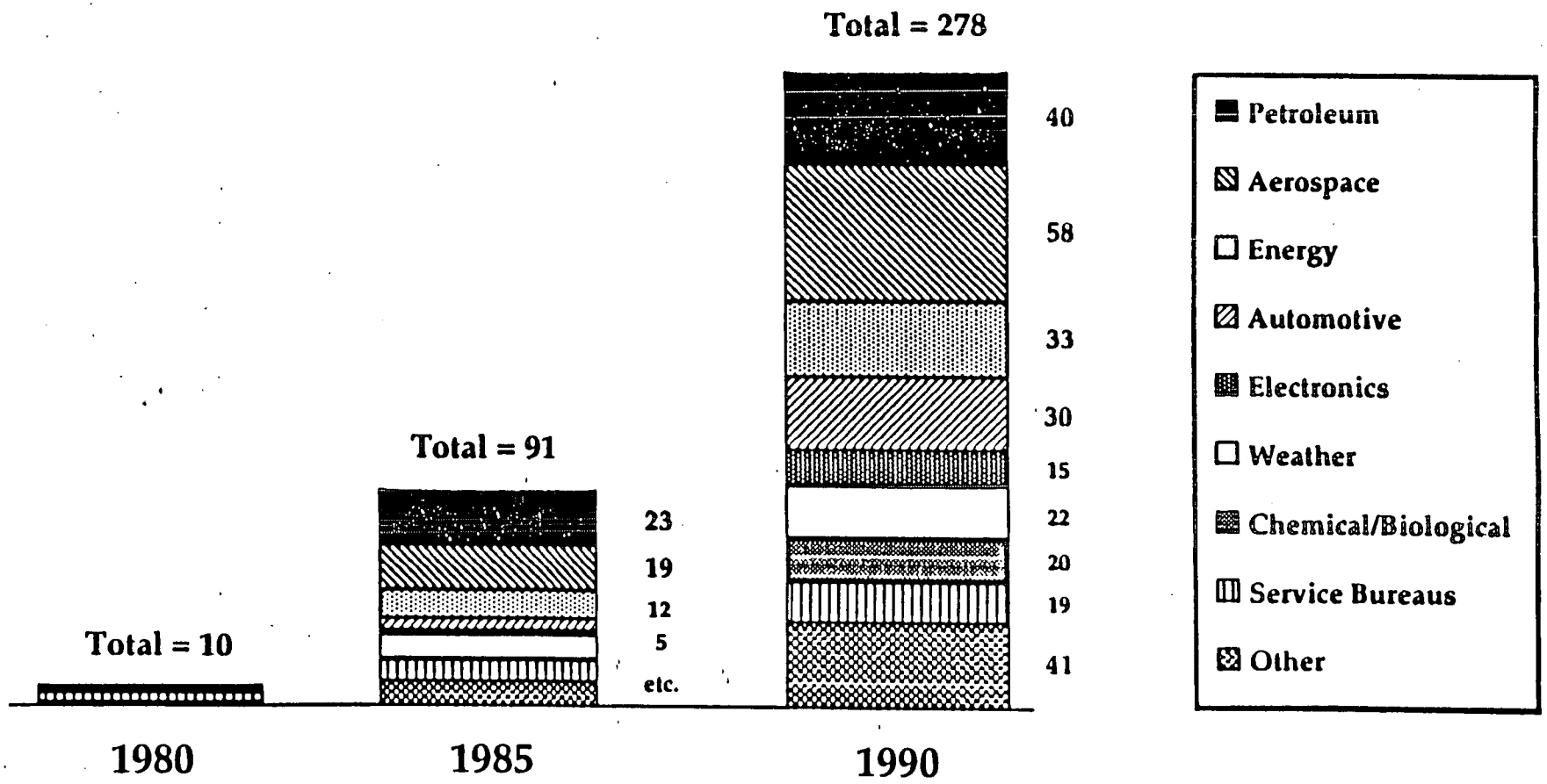
HPC is not driven by just a desire to build and use faster and bigger computer systems (although that admittedly is a motivator for some people in the field). It is driven primarily by applications which require extraordinary computing capabilities. Historically, many of these applications were defense related -- indeed, nuclear weapons design and intelligence are still major drivers of HPC development – and therefore a major portion of HPC usage took place in extreme secrecy. But over time, word of the value of HPC as a tool in scientific investigation and engineering spread into non-defense communities, and some of the tools and techniques developed for classified applications were adapted for academic and commercial settings. Thus, the pattern of HPC usage evolved over the past decade to the point where industrial usage has emerged as the dominant force in HPC markets.

The earliest industrial applications of supercomputers were in aerospace and in oil and gas exploration. Again, much of the aerospace usage was (and still is) defense related, but it is now virtually impossible to draw a line between what is military and what is civilian in most aerospace technologies. In oil and gas, the leading companies learned rather early that the use of computerized seismic exploration techniques could increase the probability of a successful well from 1 in 100 to 1 in 10. They subsequently learned that these odds could be improved still further by employing more sophisticated models. Of course, this meant still faster and more expensive supercomputers, but their cost was small compared to that of the dry wells that were avoided. (The cost of exploration has been recently estimated at $20 per barrel, as compared with about $2 per barrel for production.)

Thus, in 1985 more than a quarter of all supercomputers in industry were used in petroleum exploration. Aerospace was close behind, with just over 20 percent of the systems, and today it is the leader, with a slightly larger share. In third place, both then and now, is energy: primarily applications related to the design and operation of nuclear power plants. (See Exhibit III-2.) However, in recent years automotive usage has almost caught up with energy, thanks in no small part to aggressive investment in HPC systems by Japanese auto makers. (The total installed supercomputing power of Japanese automotive firms is four to five times that in the U.S. and double that in Europe. One Japanese company, Nissan, has as much installed supercomputing power as all three U.S. auto makers combined.)

Among the newer and more rapidly growing HPC applications are electronics, chemicals, and pharmaceuticals. In electronics, supercomputers are used, for example, in the simulation of the basic electronic devices in research and early product development and in the layout of microchips (which today may contain more than one million circuits). In chemicals and pharmaceuticals, supercomputers are an important adjunct to laboratory experimentation in determining the molecular structure and chemical/physical characteristics of new compounds for materials, drugs, etc. For example, the development of a new drug may involve investigating as many as 10,000 potential compounds, at a cost of $5,000 per compound, so if computational modeling can weed out 95 percent of the compounds, the cost of even the largest supercomputer can be saved with just one drug.

Exhibit III-2: Installed Industrial Supercomputer Systems, Worldwide



What HPC offers an industrial user is **competitive advantage**, especially in the research and development (R&D) phase of business. In particular, HPC enables:

- **More aggressive product goals;**
- **Shorter time to market; and**
- **Higher product quality.**

The quality advantage comes from using supercomputers to "design out" product defects through more detailed and exhaustive analysis than is possible with conventional computers or with laboratory experimentation and engineering prototyping. For example, in the development of the Ford Taurus, currently the best-selling American car, supercomputers were used for structural design, which reduced the amount of crash-testing necessary. This not only decreased the engineering cost and/or increased the quality achievable within the development budget, but it also enabled Ford to bring the car to market sooner, and that in turn may have allowed the Taurus to gain market share from competitors.

Certainly, both quality and early market entry are recognized factors in gaining competitive advantage in just about any market (*cf.* the writings of Harvard Business School professor Michael Porter). What is less often recognized than the quality and time factors is the ability of computers, especially high performance computers, to set and achieve more aggressive goals than would otherwise be possible (or reasonable). An example of this is given by the proposed suspension bridge over the Straits of Messina, between Italy and Sicily. If it is built, it will be the longest single-span bridge in the world, more than 2 miles (as compared with the 0.8 mile length of the Verrazano Narrows bridge in New York City), and it will have towers over 3,000 feet high and a deck more than 200 feet wide. It will take nine years to complete and will consume Italy's entire steel production for five years. The design of the bridge -- complex static and dynamic analyses, including simulations of earthquakes and other possible disasters -- would be impossible without modern supercomputers.

Supercomputers have also been used for a number of years in the entertainment industry for animation and "special effects" (especially in science fiction films), and the increasing development of visualization techniques for scientific and engineering applications may open new possibilities and markets in this area (and *vice versa*) in the 1990s. In all areas of application, the ability of advanced HPC systems to support visualization of the phenomena being investigated allows researchers to work interactively with the simulation models, thereby greatly enhancing the creative process of discovery and design.

Although comparatively little (direct) use of HPC has been made in the service sector to date, there are some discernible trends there as well. Since 1986, four large Japanese construction firms have installed (Japanese-made) supercomputers for design and structural analysis applications. Meanwhile, at least three Japanese securities companies have purchased supercomputers, and some Wall Street firms have recently acquired HPC systems for investment analysis. This use of supercomputers in the financial sector is especially significant in that it may presage the widespread use of HPC in commercial operations in the future. And when that day comes, the supercomputing stakes will go even higher.

But what is today's competitive advantage becomes tomorrow's competitive necessity, and there is a good chance that HPC may become a *sine qua non* in nearly all sectors of business in the coming years. As Alvin Toffler has explained in his most recent book, **Powershift**, there is an intensifying acceleration of all commerce taking place in the world, and this will eventually result in most business operations being conducted on a real-time basis. In such a world, the race will inevitably go to the swiftest, so using anything less than the highest performance in computing will be tantamount to conceding defeat. Unfortunately, this fact is not well understood throughout American business today, so there is reluctance on the part of most executives and managers to make the necessary investments in HPC. Faced with a multi-million-dollar minimum price for a supercomputer, many decision-makers seek a cheaper, supposedly more cost-effective, solution. But this is often "penny wise and pound foolish," as is explained in detail in Appendix F.

Grand Challenges

In addition to the industrial applications discussed above, HPC also plays a significant role in advancing all areas of science and engineering, where it serves as an essential research tool. Indeed, the use of HPC in scientific investigation has given rise to an entirely new mode of scientific inquiry, computational science, which will be explained later in this chapter.

The Office of Science and Technology Policy (OSTP) has identified 20 "Grand Challenges" in science and engineering, in which the opportunities for application of High Performance Computing are especially significant. "Grand Challenge" is defined by OSTP as follows:

> **"A Grand Challenge is 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."**

These application opportunities far surpass in importance those run on conventional mainframes, minicomputers, and desktop machines. They hold the potential for major transformations in science and engineering, product design, health, energy, security, and other vital areas of national interest.

The impacts of these application opportunities fall into the following main categories:

- National Competitiveness (including productivity, global market share, etc.)
- Energy
- Quality of Life (health, etc.)
- Basic Science
- National Security
- Others.

These application opportunities are listed in Exhibit III-3 below and categorized by "primary" or "secondary" emphasis. "Primary" areas are where we believe applications will have the strongest impact. Lesser, but significant, impacts will occur in the areas designated as "secondary."

A more detailed description of Grand Challenge and other HPC applications is given in Appendix G.

Exhibit III-3: Emphasis of Major Supercomputing Applications

APPLICATION	NATIONAL COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER
Materials Science	●			○	○	
Semiconductor Design	●			○	○	
Vehicle Dynamics	●				○	
Transportation	●				○	
Turbulence	○					●
Superconductivity	●	○	○	●	○	
Efficiency of Combustion	●	●	○			
Oil and Gas Recovery	●	●	○		○	
Nuclear Fusion	○	●	○		○	
Design of Pharmaceuticals	●		●			
Structural Biology			●	●		
Human Genome			●	○		

Legend: ● - Primary emphasis
○ - Secondary emphasis

... continued on next page

The impacts of these application opportunities fall into the following main categories:

- National Competitiveness (including productivity, global market share, etc.)
- Energy
- Quality of Life (health, etc.)
- Basic Science
- National Security
- Others.

These application opportunities are listed in Exhibit III-3 below and categorized by "primary" or "secondary" emphasis. "Primary" areas are where we believe applications will have the strongest impact. Lesser, but significant, impacts will occur in the areas designated as "secondary."

A more detailed description of Grand Challenge and other HPC applications is given in Appendix G.

Exhibit III-3: Emphasis of Major Supercomputing Applications

APPLICATION	NATIONAL COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER
Materials Science	●			○	○	
Semiconductor Design	●			○	○	
Vehicle Dynamics	●				○	
Transportation	●				○	
Turbulence	○					●
Superconductivity	●	○	○	●	○	
Efficiency of Combustion	●	●	○			
Oil and Gas Recovery	●	●	○		○	
Nuclear Fusion	○	●	○		○	
Design of Pharmaceuticals	●		●			
Structural Biology			●	●		
Human Genome			●	○		

Legend: ● - Primary emphasis
○ - Secondary emphasis

... continued on next page

Exhibit III-3 (cont'd)

APPLICATION	NATIONAL COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER
Prediction of Weather and Global Climate Change			●	○	○	
Computational Ocean Sciences			○	●		
Astronomy				●		
Quantum Chromodynamics				●		
Speech	●		○		○	
Vision	●				○	
Vehicle Signature					●	
Undersea Surveillance					●	
Engineering	●	○			○	
Computational Chemistry	●			○		
Film Animation	●					
Bond Bidding	●					

Legend: ● - Primary emphasis
○ - Secondary emphasis

The foregoing array of supercomputing applications is by no means exhaustive, but it gives an indication of the significance of HPC as well as the range of potential HPC usage. Furthermore, the historical pattern of advanced computing applications development is that performance and usage lead to new insights, opportunities, and applications. While most of the applications listed above will still be viable in the year 2000, we would be surprised if there were not a number of significant new entries on the list by then.

These application opportunities underscore the fact that High Performance Computing is a significant national priority, irrespective of increasing threats to U.S. superiority by foreign-based companies. Indeed, a good case could be made that the Federal HPCC Program would be urgently needed even in the absence of foreign competition; if only to sustain progress in science and innovation in technology.

COMPUTATIONAL SCIENCE

A common thread running throughout these Grand Challenges, and also the industrial applications of HPC described above, is computational science. Broadly speaking, computational science entails the construction of large, complex computer models of physical systems, then observing the behavior of the models under various conditions of change. For example, a meteorologist may observe the behavior of a thunderstorm model as time passes. A chemist may observe the model of a complex molecule. An automotive engineer may observe the model of a car as it crashes into (the model of) a wall. The outcomes of the insights gained by these scientists and engineers result in, for example:

- Better prediction of weather;
- Improved material science; and
- Safer automotive design and faster time to market.

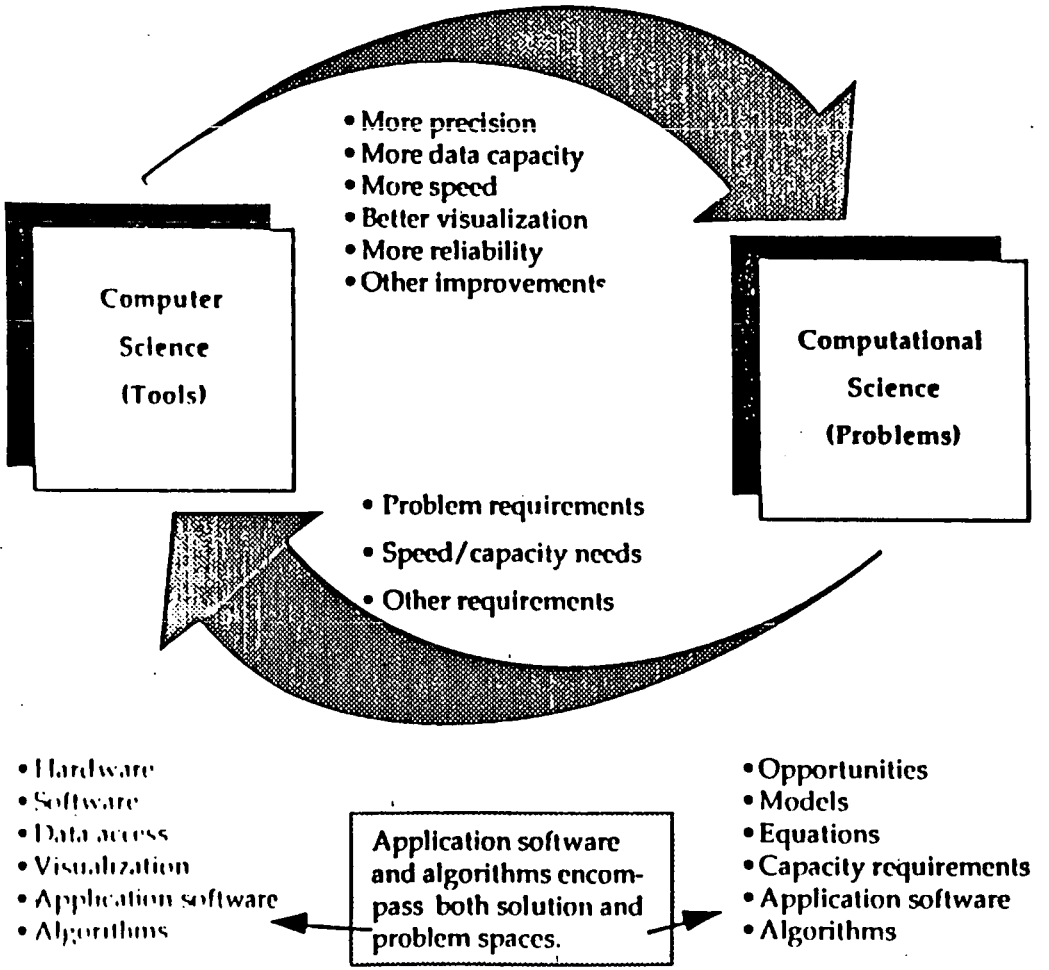
Computational science is new, and it is growing. In the last decade, its value has become established, and it has taken its place along side empirical and theoretical science as the newest scientific paradigm (see Exhibit III-4). This notion of a new paradigm for science and engineering is crucial, because it entails changes in basic outlooks and attitudes on the part of scientists, engineers, executives, and policy makers (*Cf.* Thomas S. Kuhn, The Structure of Scientific Revolutions, Univ. of Chicago Press, 1962).

Exhibit III-4: Computational Science, A New Paradigm

Type of Science	Empirical	Theoretical	Computational
Archetype	Galileo	Newton	Von Neumann
Purpose	Description	Explanation	Prediction (What if?)
Tools	Scientific Instruments	Mathematics	Computer Systems
Results	Facts	Theories	Insights through Visualization

In future years, computational science is virtually certain to be increasingly critical to national competitiveness and quality of life. But the successful application of HPC in computational science also depends upon advances across multiple fronts, including raw computing performance, software tools, visualization tools and techniques, new and improved algorithms, applications software and, not least of all, trained computational scientists in industry, academia and government. Thus, the Federal HPCC Program embraces both computer science, the craft of designing and using more powerful and useful computers, and computational science, the application of these computers to significant problems of science and industry. Computer science and computational science reinforce one another, as illustrated in Exhibit III-5. Countries that are leaders in computer science will be more capable of leadership in computational science, and *vice versa*.

Exhibit III-5: Computer Science and Computational Science



OBSTACLES

The obstacles to successful implementation of supercomputing applications include the following:

- **Computer performance** (speed and data capacity);
- **Intellectual barriers** (hard to use software, insufficient training);
- **Access barriers** (need for users to be physically close to the computer because of network limitations);
- **Algorithms** (problem solving procedures that both address applications and take advantage of new supercomputer architectures);
- **Software** (lack of problem-oriented, easy to use software -- see also "Intellectual barriers");
- **Human resources** (lack of sufficient numbers of experts in application disciplines);
- **Attitudinal barriers** (reluctance of senior executives to support the computational science paradigm);
- **Funding** (reluctance to fund supercomputing development and usage, perceived as high risk).

All of these obstacles need to be overcome if the U.S. is to seize the opportunities represented by the Grand Challenges and other supercomputing applications. The obstacles are discussed below in detail.

Performance

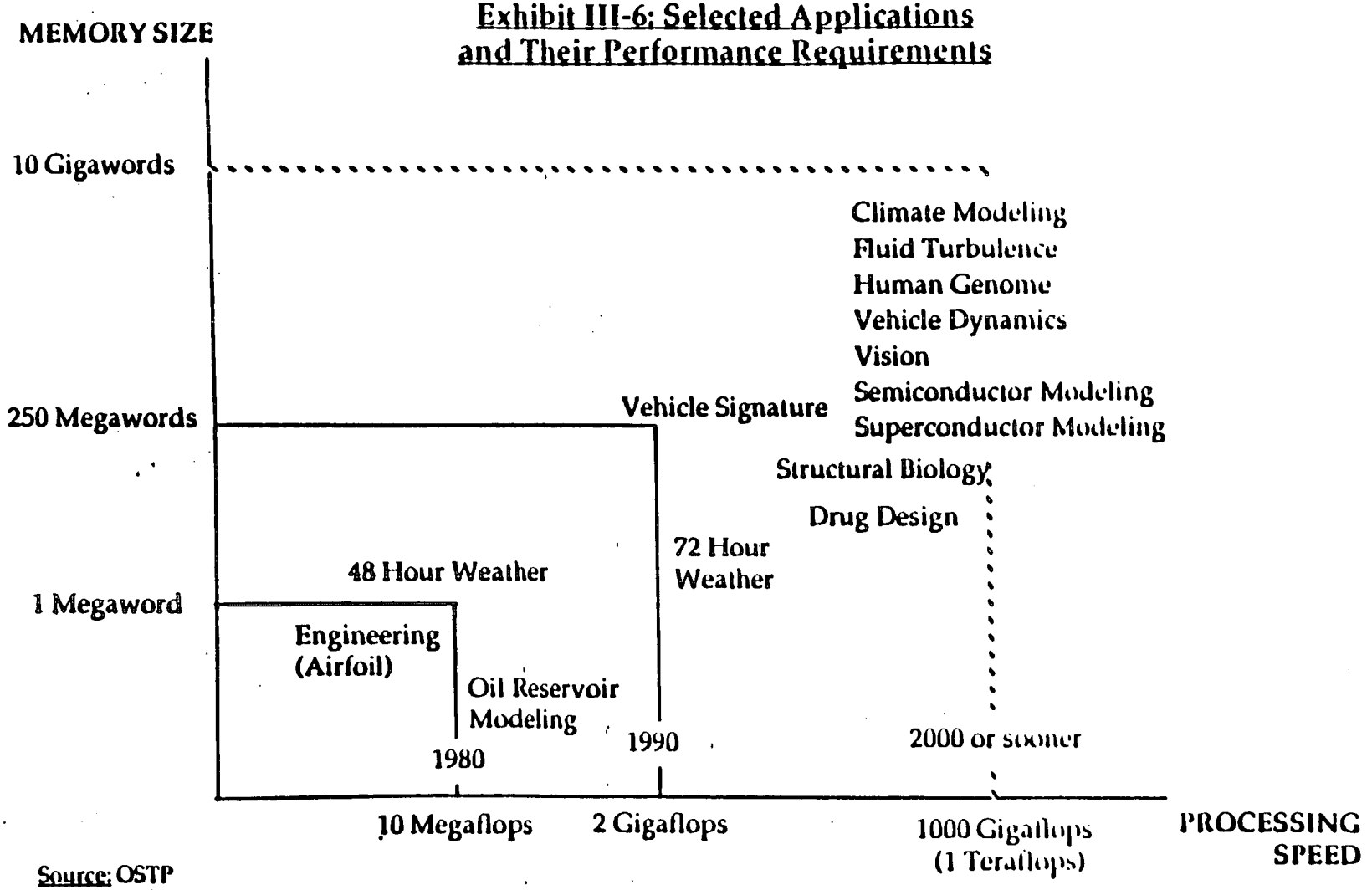
Current levels of performance on Cray supercomputers reach almost three gigaflops. A "gigaflop" stands for a billion ("giga") floating-point operations per second ("flops"). A floating-point operation is an arithmetic step such as add, subtract, multiply, or divide. The "floating" (decimal) point allows the storage of both very large numbers and very precise fractions. (See also Appendix D.)

Another gauge of supercomputer capacity is computer memory size, typically measured in megawords (a million words) or gigawords (a billion words). In supercomputers, a "word" is typically 64 binary digits ("bits") long. Memory is used to store applications and data that are needed very quickly by the central processing unit.

Exhibit III-6 shows the growth attained in the past and needed in the future in flops and word capacities of the leading supercomputers, in relation to the requirements of selected applications.

OBSTACLES

Exhibit III-6: Selected Applications and Their Performance Requirements



Source: OSTP

As shown in the preceding exhibit, many important applications require performance well beyond the capabilities of today's leading computers. Indeed, the performance barrier will cause other obstacles to come into play before it is overcome. Most computers in use today use a "Von Neumann architecture," named for the Princeton scientist who helped develop some of the earliest computers. In a Von Neumann machine, data are supplied to the computer processor one datum at a time. In this serial mode of processing, theoretically and practically attainable speeds are limited by manufacturing and packaging techniques, but most fundamentally by physical constraints such as the speed of light. The upper limit or "wall" of Von Neumann machine speed is not known precisely, but most experts would agree that it is well below 100 gigaflops.

In order to surpass the "Von Neumann wall," scientists and engineers are working on parallel architectures. In a parallel computer, multiple streams of data are supplied concurrently (in parallel) to multiple processors. For example, in a parallel machine with 100 processors, speeds approaching 100 times that of a single Von Neumann machine are theoretically attainable. The problem is that virtually all of the algorithms and computational science applications and supporting software have been developed for Von Neumann architectures. In order for parallel architectures to succeed widely, a whole new class of algorithms, applications, and software must be developed. Furthermore, computer scientists and computational scientists need to be trained and educated in parallel computing concepts and techniques.

The parallel computing challenge is a good case example of the point that obstacles to computational science are interconnected and cannot be overcome by attack on a narrow front.

Intellectual Barriers

Supercomputers are not easy to use, and computational scientists are not widely trained on them. These obstacles can be attacked successfully on at least two fronts.

First, software is needed that is functionally rich, (computational science) problem-oriented, and easy to use. Such software will allow computational scientists to minimize the amount of intellectual energy they need to divert from the real problem at hand simply in order to deal with the arcana of computing. This idea is "old hat" in the personal computer sector and needs to be pursued vigorously in the supercomputer arena. The problem is that authors and vendors of personal computing software look at potential markets in the tens of millions of customers, whereas supercomputer authors and vendors can contemplate at most a few thousand customers. Thus, market forces tend to drive the most talented software authors, who are in short supply, away from supercomputing and toward the low end of the computing spectrum.

Second, education and training are needed for computational scientists. The requirement here is not limited to computing techniques, but includes also such areas as awareness of new science and engineering approaches, parallel algorithms, and cross-disciplinary problem solving.

While the existing academic supercomputing centers and the active vendors perform creditably in education and training, a stronger effort will be needed in the 1990's if the supercomputing applications described in the preceding section are to be implemented successfully.

Access Barriers

Visualization, the graphic presentation of supercomputing models on high resolution screens, is an integral component of most of the Grand Challenge and other supercomputer applications. Existing nationwide computing networks are not capable of transferring rapidly enough the large amounts of data required for "real-time" visualization of model results. Consequently, supercomputer users often must go to computer sites, where local data channels can provide the needed capacity and response. This travel and/or residency requirement presents a disincentive to supercomputing use.

High capacity computer networks are needed to remove this barrier to physical accessibility.

Algorithms

Algorithms are formal procedures for solving computational problems. They reflect not only the characteristics of the problem to be solved, but also the architecture of the computer(s) used for the solution.

Advances in algorithms are needed for Grand Challenge and other supercomputing applications, and especially for the parallel computer architectures that will be required for performance-bounded applications.

OBSTACLES

Software

The shortage of good software has been endemic in all sectors of the computing field since its inception, and the High Performance Computing sector is no exception. Software is needed for performance and functional power and also to reduce the intellectual barriers mentioned earlier.

Particular software needs in supercomputing include:

- Languages and compilers;
- Systems (internal computer logistics) software;
- Software to manage large, distributed databases;
- Implementations of parallel algorithms;
- Visualization and debugging software; and
- Instrumentation software.

As noted earlier, the high performance software market is dwarfed in size by other sectors of the computing market and, hence, is not well served by normal marketplace forces.

Human Resources

Skilled computer scientists and engineers are needed to design and build supercomputer systems, and skilled computational scientists are needed to exploit them. Further, instructors and "teacher/teachers" are needed to provide the education and training. In the most simplistic terms, the potential of computing has always been bounded by technical capabilities on the one hand and by the human intellect on the other hand.

In the computer industry, good talent is in short supply, and most observers predict that the shortage will become even more severe in the 1990's (see Appendix H). The Grand Challenges and other supercomputing applications require strong computational scientists to push the envelope of existing systems and equally strong computer scientists and engineers who can deliver improved systems.

Attitudinal Barriers

Supercomputing installations are expensive. Therefore, their acquisition and maintenance budgets require approval by senior executives who often have little knowledge or experience in computational science. Computational science is a new paradigm, with which many senior executives will feel uncomfortable. While their skeptical attitudes are justified in part, they are also dysfunctional to the extent they are based on inertia, ignorance and fear of change. In the automotive industry, for example, many executives will be more comfortable with real prototype cars crashing into real brick walls than with a computer model of the same event, even though the model will facilitate better and faster design. This attitudinal barrier may fade away in 20 or 30 years as computer-trained generations reach executive ranks, but by then today's competitive wars will be history and the losers will be all but forgotten.

Executive-level awareness-building and education are needed to overcome the attitudinal obstacles. Domestic competitive success stories will also do a lot to change executive attitudes.

OBSTACLES

Funding

The expense of a supercomputer installation cannot be borne easily, especially by academic institutions. On the supply (vendor) side of the private sector, supercomputer research and development may get underfunded because of the risks involved, the presently small size of the supercomputing market, and competing R&D projects with payoffs that are judged to be nearer term and/or more certain.

Government funding support is needed to promote academic supercomputing and to reduce the risks of private sector supercomputer development and usage.

In terms of these obstacles, the Federal HPCC Program is targeted as shown in Exhibit III-7.

Exhibit III-7: Impact of HPCC Program on Obstacles to Supercomputing

	COMPONENTS OF HPCC PROGRAM				
Obstacles	HPC Systems	Software & Algorithms	NREN	Research & Human Resources	Comments
Performance	●	●		○	Hardware and architecture
Intellectual		●	○	●	Training and software ease of use
Access			●	○	Network extends physical access
Algorithms		●		○	Needed for problems and for parallel architectures
Software		●		○	Can never be rich enough
Human Resources		○	○	●	The other side of technology
Attitudes	○	○	○	○	Special form of education
Funding	○	○	○	○	HPCC helps across the board

● Primary thrusts of HPCC component
○ Secondary thrusts of HPCC component

[This page has been left blank intentionally.]

CHAPTER IV - THE HPC ARENA

IV - THE HPC ARENA

This chapter provides an overview of the present situation in HPC. The chapter consists of five sections, as follows:

- **The Past Decade** - traces the evolution of the supercomputer market from 1980 through the present.
- **The Next Decade** - explains the framework used in developing two alternative scenarios for the supercomputer industry through the year 2000.
- **Scenario A** - presents our projection of the supercomputer industry under the assumption that the Federal HPCC Program is not funded.
- **Scenario B** - presents our projection of the supercomputer industry under the assumption that the Federal HPCC Program is funded.
- **HPCC Program Impact** - describes the projected impact of the Federal HPCC Program on the supercomputer industry, 1990-2000, by comparing and contrasting Scenarios A and B.

THE PAST DECADE

The 1980s have witnessed a flowering of High Performance Computing into a full-fledged segment of the information industry. At the beginning of the decade, there were only two supercomputer vendors, Control Data Corporation (CDC) and Cray Research, both American firms and both following a vector processing approach to supercomputing performance (see Appendix D). In 1983, the situation changed abruptly as the three largest Japanese computer companies, Fujitsu, Hitachi, and NEC, all announced supercomputer systems. The shock of this action was compounded by the fact that the performance ratings of the Japanese machines, ranging up to 1.3 gigaflops, surpassed (at least on paper) that of the best American-made systems at that time. However, Cray Research countered with a four-processor version of its X-MP in 1984 and the 2 gigaflops Cray-2 (also a four-processor system) in 1985, which kept the performance championship (as measured by theoretical peak megaflops) in American hands, at least for the time being. CDC also responded by announcing the formation of a subsidiary, ETA Systems, to develop and build a new line of supercomputers with peak performance up to 10 gigaflops.

Meanwhile, several companies, following the lead of Floating Point Systems (FPS), began marketing "minisupercomputers," which offered near-supercomputer performance at a fraction of supercomputer prices, and soon a number of new firms -- such as Active Memory Technologies, Alliant, BBN, Convex, FPS, MasPar, Meiko, and nCUBE -- were fighting for survival in this highly competitive market segment. In the mid-1980s, IBM entered the fray by adding a Vector Facility to its 3090 mainframe systems to endow them with performance approaching supercomputer levels, and DEC did likewise at the end of the decade when it announced its VAX 9000 line of mainframe systems. The Japanese vendors also introduced scaled-down versions of their supercomputers to compete against the minisupers, and Cray Research moved to reduce its exposure at the low end of the market by making available some entry-level systems with reduced prices (and capabilities).

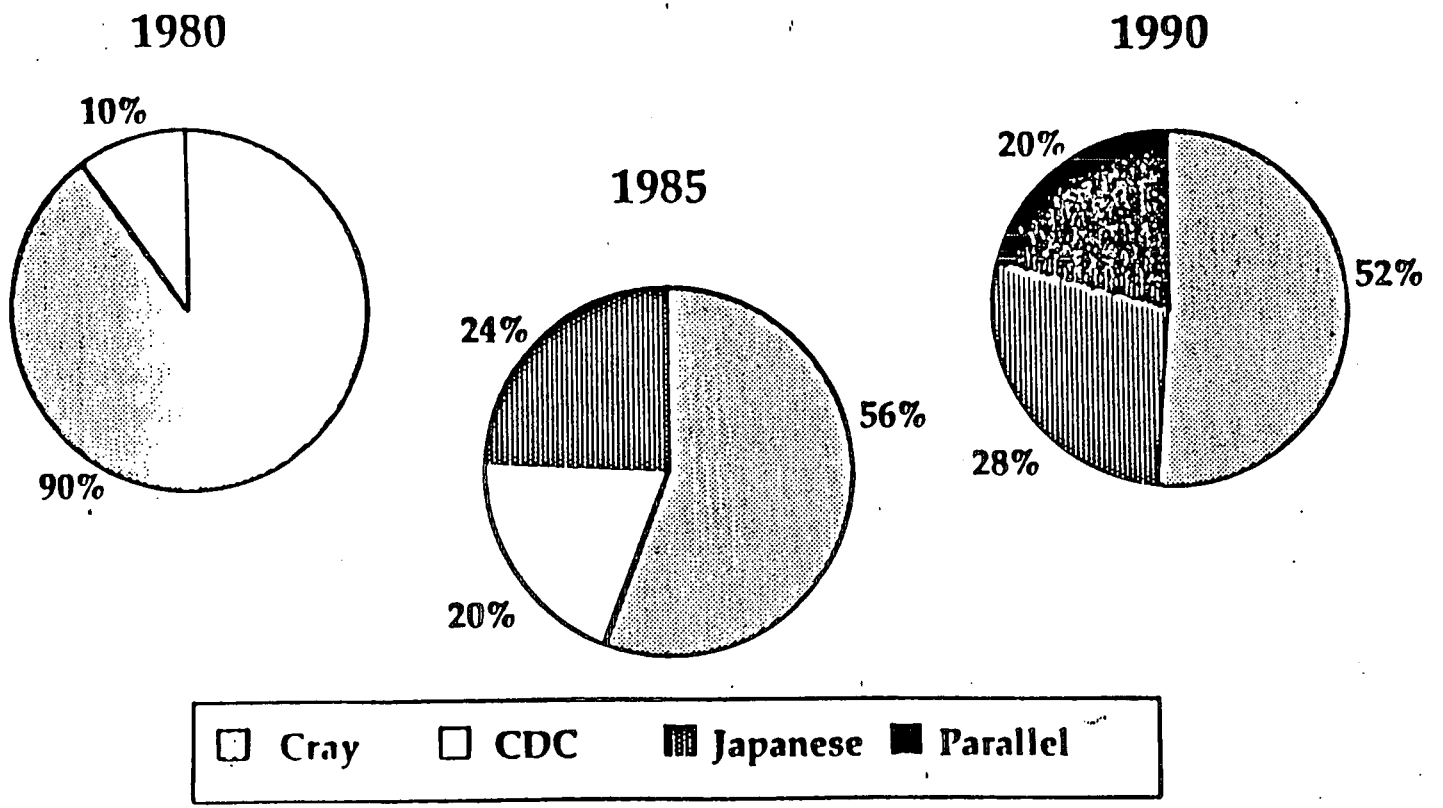
In 1988, the Japanese began announcing their "second generation" of supercomputers: the Hitachi S-820, with performance up to 3 gigaflops; the Fujitsu VP-2000 series, with performance up to 4 gigaflops (since upgraded to 5); and the first Japanese multiprocessor supercomputer, the NEC SX-3 (marketed in the U.S. as the SX-X), with peak speed in excess of 22 gigaflops. Almost immediately, U.S. experts expressed doubt as to whether these theoretical peak speeds could be approached in everyday usage, both because of the extreme architectural characteristics of the hardware and also because of the primitive level of the software (see Appendix D). Nevertheless, it was widely perceived that the Japanese were raising the ante in the supercomputing game by using their prowess in semiconductor technology and that U.S. supremacy in overall supercomputer functionality and performance is likely to come under increasing challenge from Japan in the future. The sense of crisis was heightened by Control Data's sudden decision to shut down its ETA Systems venture and withdraw from the supercomputer business in early 1989, and by the subsequent spin-off of Cray Computer (along with founder and chief designer Seymour Cray) from Cray Research.

At the same time, a new challenge to established supercomputing orthodoxy was emerging from the ranks of minisupercomputer makers: large-scale and massive parallelism (see Appendix D). Although most vendors following this approach concentrated on systems competing with minisupers in price and performance, two of them, Intel and Thinking Machines, began to offer systems with peak theoretical performance in the tens of gigaflops. Again, many experts refused to regard these new competitors as genuine threats to Cray, using essentially the same arguments levied against the Japanese: the architecture is extremely specialized and the software is relatively crude, both of which limit the performance actually attainable in all but a few isolated situations. However, Thinking Machines and (to a lesser extent) Intel continued to rack up impressive achievements in an ever-widening range of applications and have thus gained grudging acceptance into the ranks of supercomputer makers. This was confirmed by Cray Research's October, 1990, disclosure that it was establishing a major development effort centering around the highly-parallel approach and by announcements from Japanese vendors, notably Fujitsu and new entrant Matsushita, that they too would soon bring highly-parallel systems to market.

Although European firms remained on the sidelines in the supercomputer competition of the early 1980s, they also began to show signs of interest in highly-parallel systems, largely based upon the Inmos "transputer" chip. IBM also intensifyed its presence in all aspects of HPC through a number of actions: investment in Supercomputer Systems, Inc., a company formed by former Cray Research designer Steve Chen to develop a parallel supercomputer; continuing enhancement of its mainframe Vector Facility, with a significantly more powerful, redesigned unit for the top-of-the-line models 820 and 900 in the new ESA/9000 series; downward expansion of its vector processing offering into the air-cooled (9121) portion of its ESA/9000 line; and aggressive development and marketing of its RISC-based RS/6000 line of HPC workstations and departmental systems. Although some experts did not consider IBM to be a full-fledged supercomputer maker in the 1980s, there are definite indications that it will become one in the not-too-distant future, if not through one (or more) of the aforementioned initiatives, then perhaps through derivatives of its internally-developed, highly-parallel research prototypes, the GF11 and RP3.

Thus, the pattern of supercomputer system shipments has changed substantially over the past decade. As shown below in Exhibit IV-1, the dominance of Cray Research in the industry has eroded significantly -- although it must be cautioned that these figures are somewhat biased, because small systems such as the Fujitsu VP-30E (rated at 220 peak megaflops) are counted equally with the largest Cray Y-MP (rated at 2,667 peak megaflops). (Instead of using number of systems shipped, a breakdown of peak megaflops shipments would present a somewhat different picture, albeit distorted in a different way, because Japanese supercomputers and highly-parallel systems have yet to demonstrate in general usage the levels of performance indicated by their peak megaflops ratings. See Appendix D.)

Exhibit IV-1; Supercomputer Systems Shipments



The upsurge in Japanese competition in supercomputers is part of an overall pattern of increasing Japanese presence in all worldwide information systems markets. Exhibit IV-2 provides some evidence of this trend, while Appendix I presents a more detailed discussion of Japanese activities in information technology in general and HPC in particular.

Exhibit IV-2: The Power Shift in the Worldwide Information Industry

	Number of Datamation 100 Companies		Total Information Systems Revenue		Revenue CAGR
	1983	1987	1983	1987	
United States	71	60	$87 B	$132 B	10.8%
Europe	19	22	$12 B	$35 B	17.5%
Japan	8	16	$9 B	$40 B	27.9%

Source: The Competitive Status of the U.S. Electronics Sector: From Materials to Systems; a report of the Secretary of Commerce to the House Appropriations Committee; December, 1989.

Worldwide revenue growth in supercomputers (excluding IBM) has been nearly double that of the information industry as whole (see Exhibit IV-3), but as compared with the rest of the industry, supercomputer revenues are still relatively minuscule: Worldwide sales have barely exceeded $1 billion in recent years (see Exhibit IV-4.) That number would be perhaps $400 million higher if IBM's supercomputer revenues* are included, and perhaps again that much greater if minisupercomputers were also included.

* In IBM's case, it is very difficult to determine the "correct" revenue figure: Should it be only the income derived from the Vector Facility units -- about $190,000 each -- or should it be the entire amount from the systems -- ranging up to more than $20 million each -- of which they are a part, or somewhere in between? IBM does not disclose an "official" figure for revenues, number of VF units, or systems.

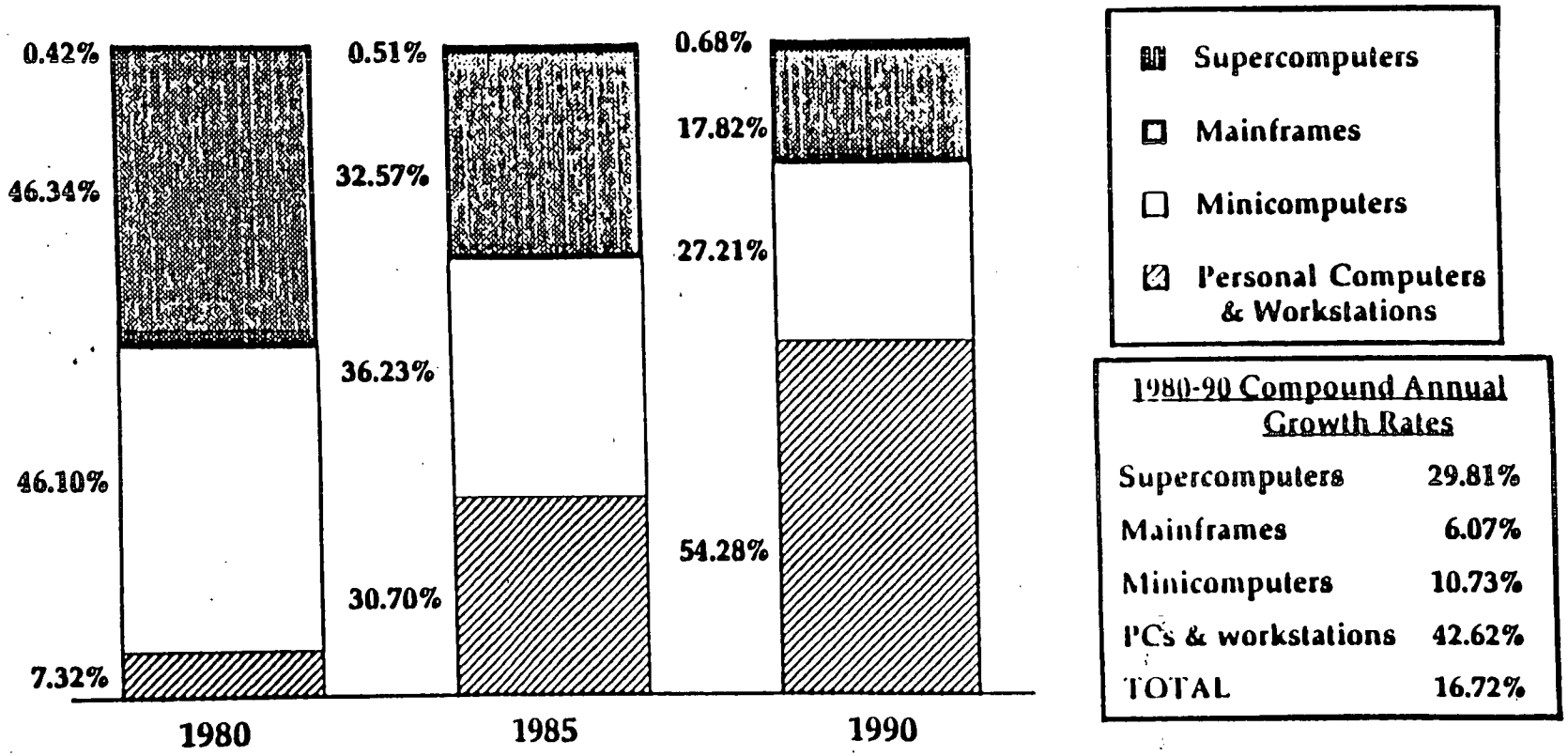
In addition to expanding supercomputer usage in the U.S., the HPCC Program might also lead to increased usage of U.S.-made supercomputers overseas, but we have not attempted to estimate that in our Scenario B projections. It is also possible that U.S. consumption of Japanese-made supercomputers would be increased as a result of heightened interest in HPC. On the other hand, this might be offset by increased U.S. exports to Japan as well, so we have made no changes in our projections of Japanese supercomputer production relative to Scenario A.

The HPCC Program should also improve supercomputer price/performance, albeit only slightly, above that projected in Scenario A. For U.S.-made vector supercomputers, this would come through moderately increased demand and the consequent greater economies of scale in production. For parallel systems, price/performance improvements would be the direct result of increased R&D, plus significantly greater demand.

> Assumption #4: We assume that the price/performance of U.S.-made vector supercomputers will improve at one percentage point faster than the rates used in Scenario A. (This is a deliberately conservative assumption.) For parallel supercomputers, price/performance improvement will gradually approach levels typical of microprocessor chip technology (*i.e.*, 30+ percent per year) by the year 2000.

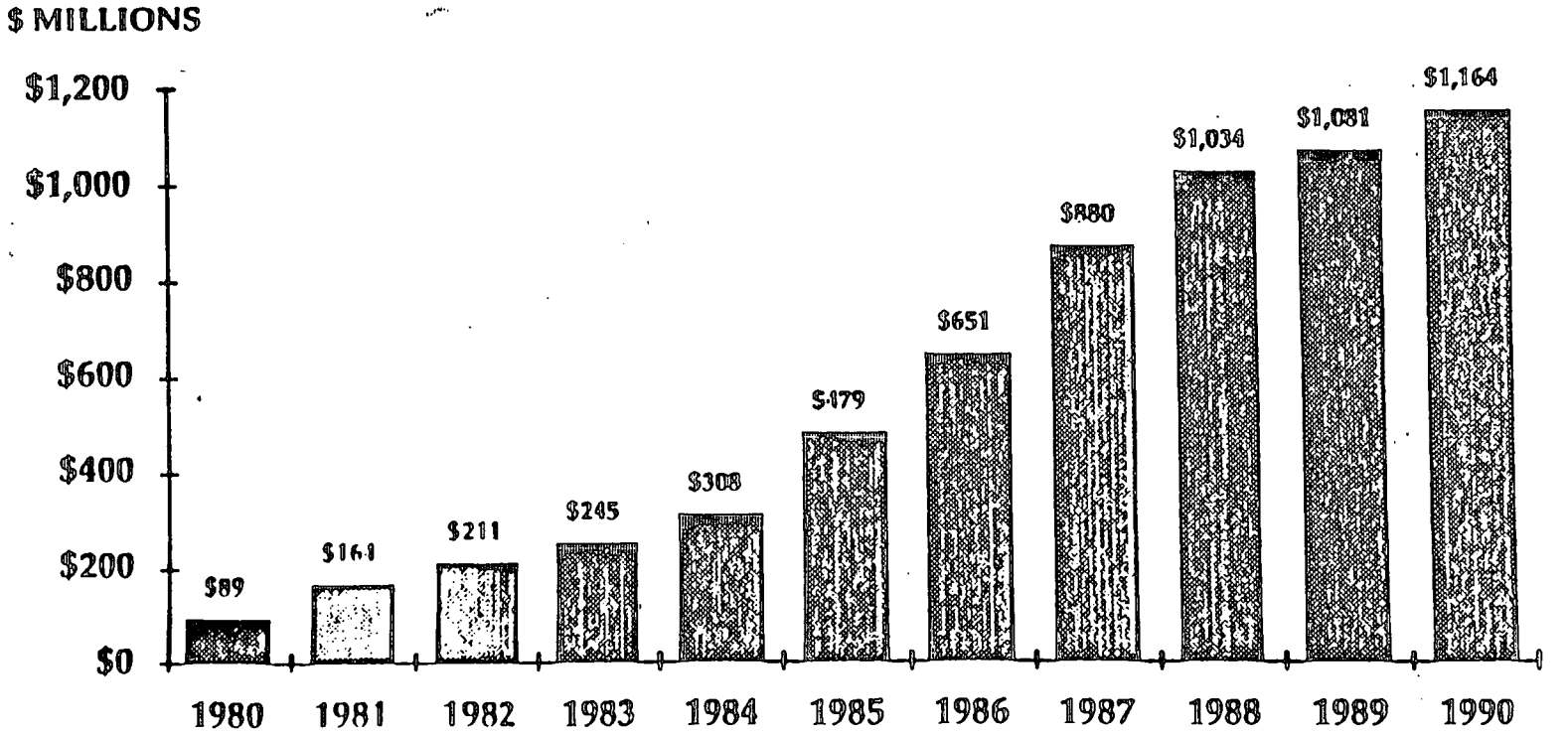
The increased R&D stimulated by the HPCC Program should also result in significantly more powerful parallel supercomputers: for example, a teraflops (1,000 gigaflops) system by about 1996. On the other hand, we do not assume any change in processing power for vector supercomputers, as compared with Scenario A, because we expect that the HPCC Program will have little effect upon hardware development for such systems. (This is distinct, however, from R&D into the use of and algorithms for vector systems, which definitely will be addressed by the HPCC Program.)

Exhibit IV-3: Value of Worldwide Computer Shipments



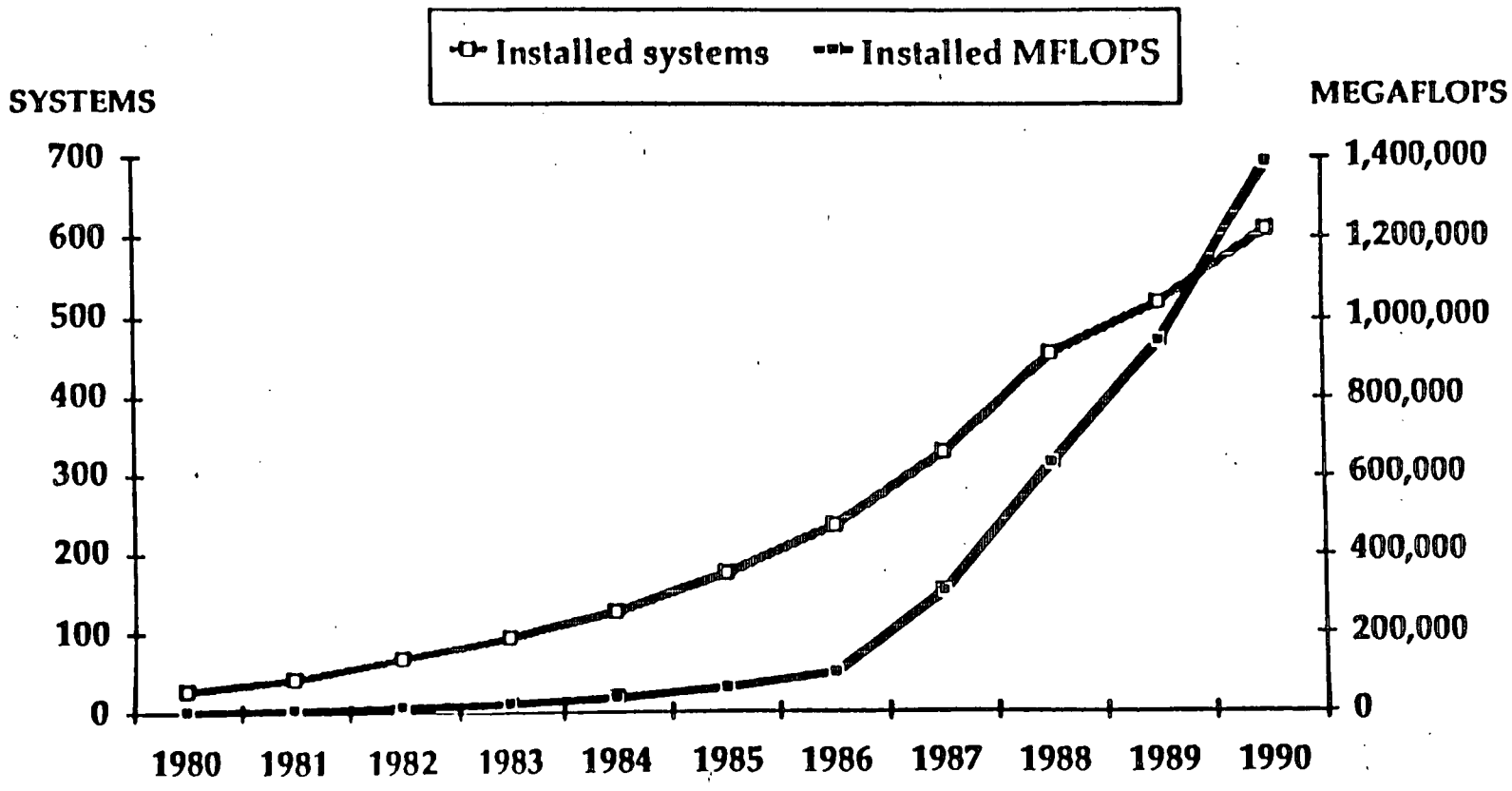
1980-90 Compound Annual Growth Rates	
Supercomputers	29.81%
Mainframes	6.07%
Minicomputers	10.73%
PCs & workstations	42.62%
TOTAL	16.72%

Exhibit IV-4: Value of Worldwide Supercomputer Shipments



As Exhibit IV-5 shows, there has been a general slowing of demand for new supercomputer systems in the past two years. Some see this merely as the result of reduced defense procurements, while others feel that it indicates a maturing industry. We tend to see it as a reflection of the "Obstacles" discussed in the previous chapter, particularly as they pertain to industrial usage of supercomputing. Our analysis suggests that the market may indeed be approaching saturation in terms of the number of potential supercomputer users and applications -- at least until such time as new markets can be opened through initiatives such as the proposed Federal HPCC Program -- but that existing supercomputer users have a virtually limitless appetite, constrained only by budgets, for computing power. In terms of megaflops, the demand for supercomputers has actually intensified in recent years.

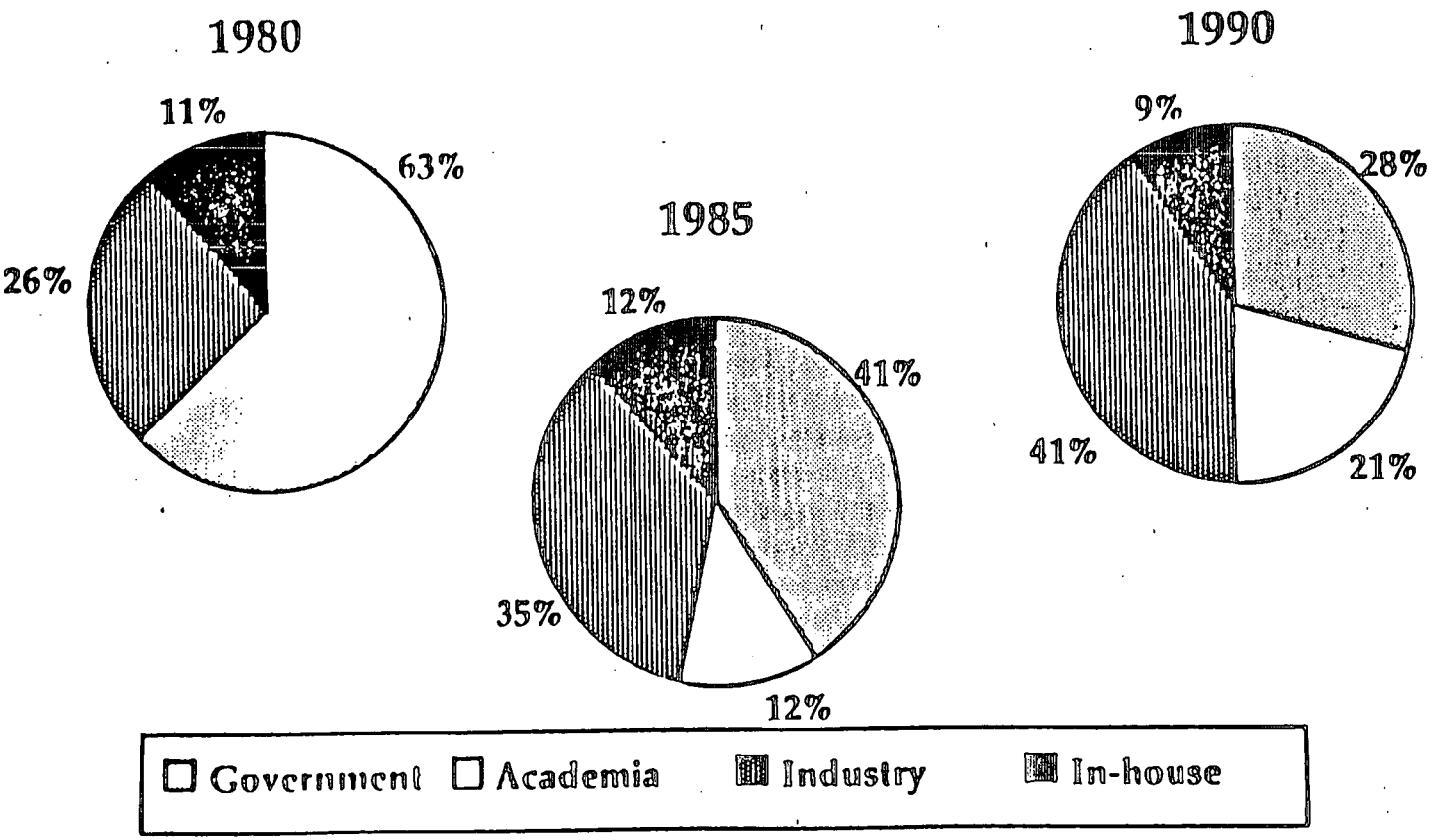
Exhibit IV-5: Worldwide Supercomputer Installations



So why, then, have the three largest Japanese computer firms invested such considerable amounts of money -- we estimate $100 million each -- to get into such a small and highly competitive market? And why, after a fifteen-year hiatus, is IBM returning to the supercomputer arena? In the case of the Japanese, there is a significant element of prestige involved. But the Japanese are not inclined to invest their money foolishly, so there must be a deeper reason. That is, we think, the very strong desire not to be dependent upon any other nation, even a "friend" such as the United States, for something so important to their long-term scientific and technological strength. The Japanese have perceived correctly the value of supercomputers in building and maintaining their industrial competitiveness, especially in "high-tech" areas: for example, electronics, biotechnology, and new materials.

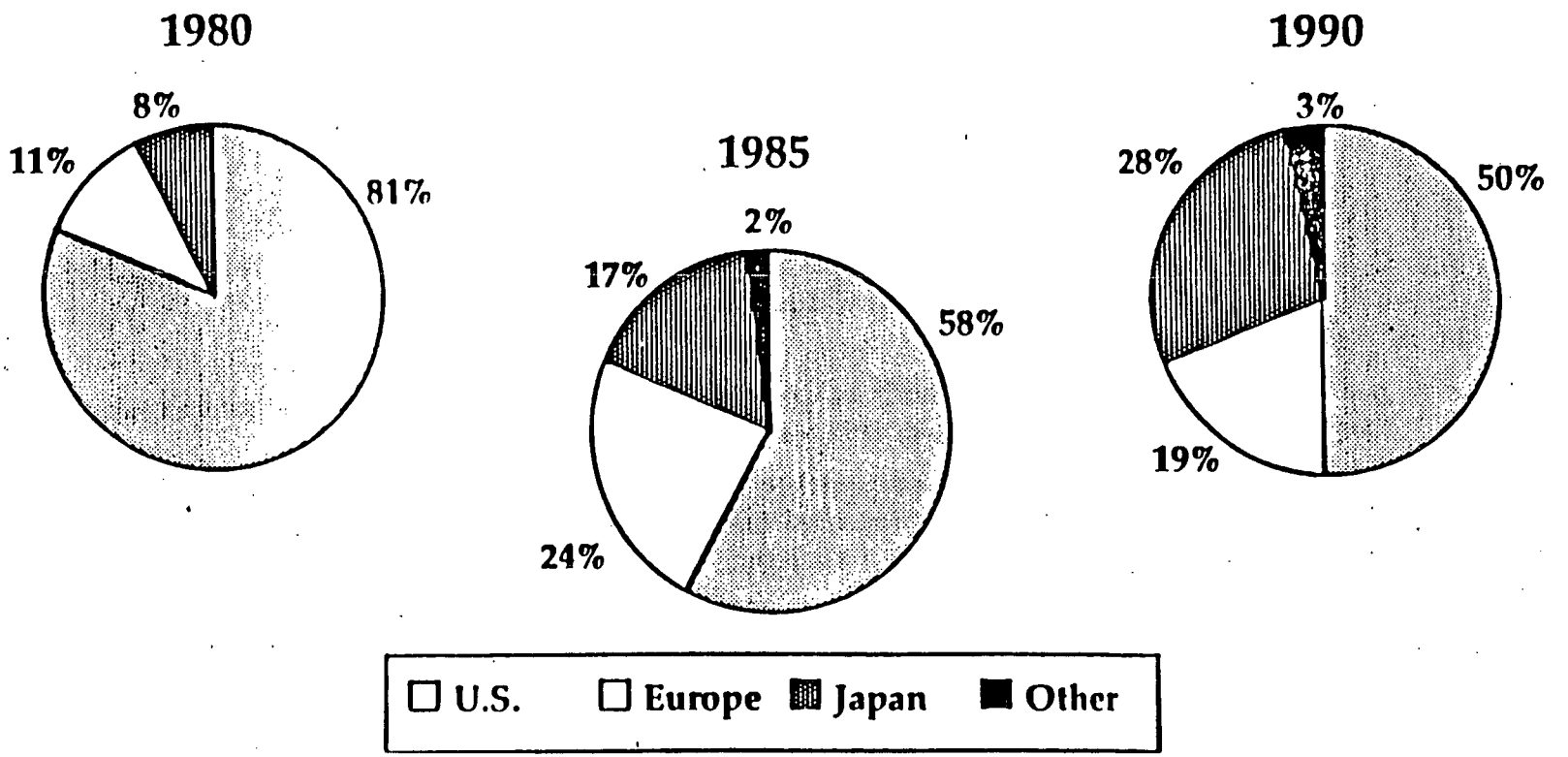
This premise is supported by the data on worldwide supercomputer installations, which show the striking change that has taken place in the supercomputer market in the 1980s (see Exhibit IV-6). Whereas nearly two-thirds of the installed supercomputers were in government laboratories in 1980, the government share is now less than half as much. This is not to say that government supercomputer installations have decreased in actual numbers, but government laboratories are certainly not the dominant customers they once were. Academic usage, much of which is also ultimately dependent upon government support, has increased from nil in 1980 to a significant share of the total in 1990. But it is industrial applications that have taken over the lion's share of the market: if the "in-house" systems retained by supercomputer companies for R&D and marketing are included, private industry now has about half of the supercomputer systems installed worldwide. (In Japan, industrial usage accounts for about 65 percent of the supercomputer installations, and in light of the anticipated payoff from the application of High Performance Computing -- see Chapter V -- the implications for Japanese competitiveness in the coming decade are indeed ominous for the U.S. and Europe.)

Exhibit IV-6: Installed Supercomputer Systems, by User



In addition to the emergence of private industry as the dominant supercomputer user and the more recent trend toward parallel systems, an even more pronounced shift in worldwide supercomputer markets has occurred in terms of geographic distribution (see Exhibit IV-7). Whereas the U.S. accounted for more than 80 percent of all installed supercomputers just a decade ago, it now has barely more than half of them. Thus, although the number of systems in the U.S. has grown twelve-fold (to approximately 300 by year-end 1990), European installations have jumped from 3 to 110 in the same period and Japanese systems from 2 to more than 160. One major reason for the latter's growth is, of course, the entry of Japanese vendors into the supercomputer market, but the strength of Japanese demand is also noteworthy: their rate of installation growth (over 56 percent) over the past decade has been nearly double that in the U.S. (less than 30 percent). It is true that many of these systems are entry-level models, closer in computing power to a minisupercomputer than a Cray Y-MP8, but it also shows the marketing acumen of the Japanese vendors. These entry-level models, priced about the same as a large mainframe and capable of running under Japanese mainframe operating systems, may have helped entice otherwise reluctant Japanese users into the supercomputer camp, a market penetration strategy which Cray Research has recently begun to emulate.

Exhibit IV-7: Installed Supercomputer Systems, by Country



THE NEXT DECADE

To assess the probable impact of the proposed Federal High Performance Computing and Communications Program, we have formulated two "scenarios" (see Appendix B for a discussion of scenarios and the methodology used in their development) for the 1990s:

- **In Scenario A**, we assume "business as usual," without any additional Federal support for HPCC beyond those activities which are now underway; and
- **In Scenario B**, we assume that the Federal HPCC Program, as proposed by the Office of Science and Technology Policy (OSTP), will receive full funding and support -- that is, $1.917 billion additional over a five-year period (FY 1992-1996).

Thus, the difference between these scenarios will be the impact of Federal HPCC Program.

In a broad sense, we expect that the principal effects of the Federal HPCC Program will be threefold:

- First, it will affect at least some of the directions of change in HPC;
- Second, it will affect the rate of change in HPC; and
- Third (and probably most important), it will affect the rate of application of HPC throughout American industry, academia, and government.

The first two of these will have a direct impact upon HPC vendors and researchers. That impact is delineated in the two scenarios for the supercomputer industry which are presented in the following two sections. The third will directly affect HPC users, as described in the next chapter.

SCENARIO A

We have framed our scenarios in terms of only supercomputers. This is not to deny the participation of minisupercomputers, workstations, networks, software, etc. in HPC, but supercomputers are the *sine qua non* of HPC. They are where the HPC "trickle down" begins and, hence, are the primary focus of the Federal HPCC Program.

Our scenarios are based upon the technological options available to makers of supercomputers in the 1990s. An alternative approach would be to focus upon vendors, but this would not provide the needed insights into how supercomputing is being done and it would tend to focus attention inappropriately on "winners and losers." The technologies available will delimit performance and pricing levels, which are the principal determinants of what is possible and what is practical in HPC.

It is, of course, impossible to predict with certainty which technologies will dominate HPC at the end of this decade, but the experience of the past decade can provide some indications. Although vector architectures have monopolized HPC through most of the 1980s, parallelism, which (except for ILLIAC IV) was virtually non-existent in 1980, seems very likely to play a major role in the future. Cray Research started down the parallel path when it introduced the two-processor X-MP in 1983, followed by a four-processor version in 1984. In 1988, an eight-processor Y-MP was announced, along with extensions to the compilers and operating system software to greatly enhance users' ability to use all processors simultaneously on a single problem. (In the X-MP, most users were limited to a single processor; multiple processors enabled multiple users to run concurrently.) Systems with still larger numbers of processors are now in development at both Cray companies: for example, the Cray-3 at Cray Computer and the Y-MP16 (a.k.a. the C-90) at Cray Research are to utilize 16 processors.

At some future point, perhaps when the number of processors exceeds 16, systems such as these will cross over the exceedingly fine line between "vector" supercomputers and "parallel" supercomputers. That is to say, they will reach the point where parallelization, not vectorization, becomes the primary means of raising performance above what can be attained by a single scalar processor. There they will be met by systems, such as those from Intel and Thinking Machines, which have employed large amounts of parallelism from the very outset and which may or may not have added vector processing along the way.

This is not to say that vector supercomputers will go the way of the dinosaur. They will continue be used well into the next century. In a sense, the situation between vector and parallel supercomputers in the coming decade will be like that which has been playing out between mainframes and desktop systems since the IBM PC was announced in 1981. Although the desktop systems (and, analogously, highly-parallel supercomputers) appear to offer much better price/performance than mainframes (or vector supercomputers), there are some applications which simply will not fit on the former with today's state of the art. And although this will indubitably change over time, the cost of converting other applications which run quite well on mainframes (or vector supercomputers) may never be justified. Hence, these two classes of applications will continue to drive demand for mainframes (and vector supercomputers) even though desktop systems (and highly-parallel supercomputers) become the preferred platforms for new applications in the 1990s.

For present purposes, it is not necessary for us to distinguish among the many varieties of parallelism (see Appendix D), other than to separate (predominantly) vector supercomputers from (predominantly) parallel systems because of significant differences in their overall characteristics (see Exhibit IV-8). Having done that, it is also necessary to distinguish Japanese-made vector supercomputers from U.S. made vector supercomputers for the same reason.

Exhibit IV-8: Typical 1990 Supercomputer Characteristics

	Peak Megaflops*	System Price	$/Peak Megaflops*
U.S. vector supercomputers	1,337	$11.8M	$8.8K
Japanese vector supercomputers	3,565	$10.4M	$2.9K
Parallel supercomputers	10,500	$3.6M	$0.34K

At some future time, the characteristics of the various types of parallel systems, including their country of origin, may diverge enough to justify subdivision of the "parallel supercomputer" class. For the present, however, Japanese and European involvement in this sub-market is virtually non-existent, and we can distinguish no other reason for differentiating among the various systems available in the analysis which follows.

* Important Note:

As before, we caution against misinterpretation of the figures in Exhibit IV-8 and in any and all of the following Scenario A and Scenario B Exhibits which are based upon peak megaflops ratings.

Peak megaflops ratings tend to be extremely misleading, because the ratio of peak megaflops to attained performance in an actual application varies greatly among these three classes of supercomputers (and also from application to application).

Peak megaflops ratings are used in this report only because (a) they are readily available for all current supercomputer systems and (b) there is no universally accepted alternative measure of supercomputer processing speed.

We now begin our elaboration of Scenario A.

Assumption #1: We assume that the developers and builders of supercomputer systems in the next decade can be grouped as follows:

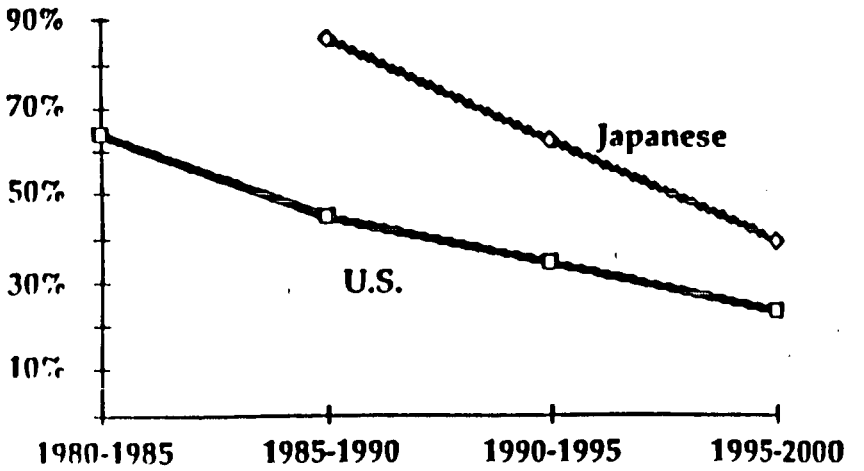
- **U.S. vector supercomputer vendors** -- Cray Research, Cray Computer, and (for historical purposes) Control Data Corporation (and its supercomputer subsidiary, ETA Systems), and perhaps IBM and Convex;
- **Japanese vector supercomputer vendors** -- Fujitsu, Hitachi, and NEC; and
- **Parallel supercomputer vendors** -- current market participants such as Intel, MasPar, nCUBE, and Thinking Machines, plus anticipated new entrants, including IBM, Cray Research, Convex, and Japanese and European companies.

We do not anticipate any future European participation in the vector supercomputer market; indeed, European vendors have all but disappeared from the mainframe business as well, except as remarketers of U.S.- and Japanese-architected systems.

Supercomputer usage has shown dramatic growth in the 1980s, first in U.S.-made vector supercomputers, then in their Japanese-made counterparts, and most recently in parallel systems. However, growth rates typically decrease over time, although the rate of decrease can be cut by government initiatives, such as the proposed Federal HPCC Program, which stimulate innovation and new applications.

Assumption #2: We assume that the signs of maturity which have been observed in the market for vector supercomputers since 1988 will become even more evident in the latter 1990s, after the current generation of Japanese supercomputers and the next generation of U.S. vector supercomputers, the C-90 from Cray Research and the Cray-3 (and 4?) from Cray Computer, have had their day (see Exhibit IV-9).

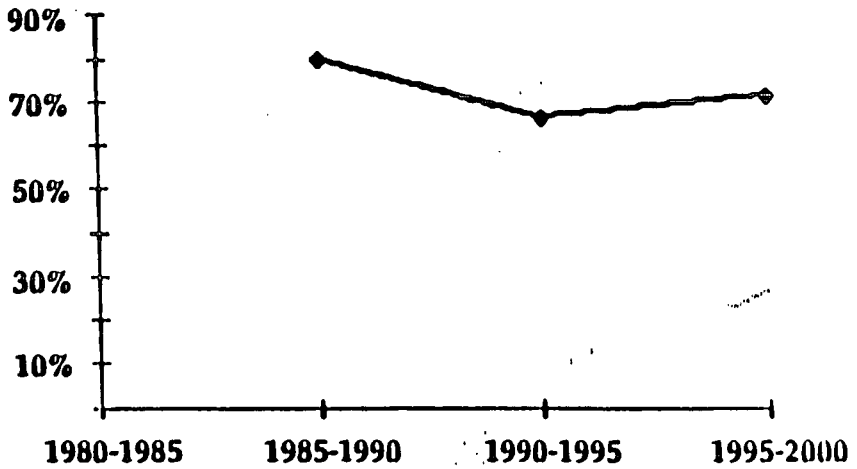
Exhibit IV-9: Growth Rates for Vector Supercomputers, Scenario A
(installed peak megaflops)



Apart from market maturity, we expect that sales of vector supercomputers will decline in the late 1990s because of the increasing popularity of parallel supercomputer systems which offer superior price/performance and overall performance -- the largest parallel systems will exceed 1 teraflops (1,000 gigaflops) by the year 2000.

Assumption #3: We assume that the recent success of parallel supercomputers in certain applications will expand to other areas once the technical difficulties with programming and algorithms are overcome. When that happens, usage of parallel systems will increase significantly (see Exhibit IV-10).

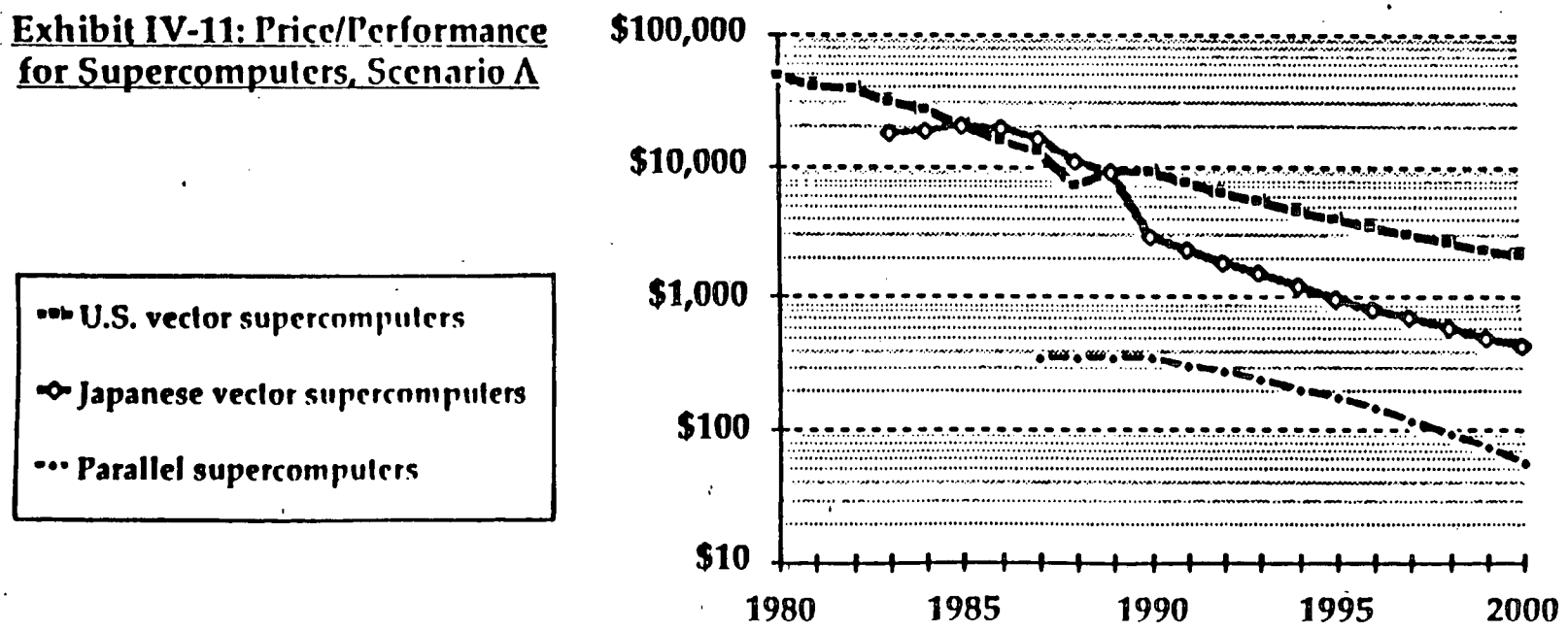
Exhibit IV-10: Growth Rates for Parallel Supercomputers, Scenario A
(installed peak megaflops)



Since 1980, the price/performance (as measured in $/peak megaflops) of U.S.-made vector supercomputers has improved -- that is, decreased -- at an average rate of just over 15 percent per year. Over a shorter period, Japanese-made vector supercomputers have shown more than 30 percent average annual price/performance drops, much of that being the result of the recent introduction of "second generation" Japanese supercomputers. The price/performance of parallel systems has not changed much in the few years since their commercial introduction, but it is much lower than that of vector systems to begin with. In the coming years, it will also begin to fall.

Assumption #4: We assume that the price/performance of vector supercomputers will continue to improve, albeit at slower rates after 1995 because of a shift in R&D focus toward parallel systems. We also assume that the price/performance of parallel systems will improve at an accelerated rate after 1995 (see Exhibit IV-11).

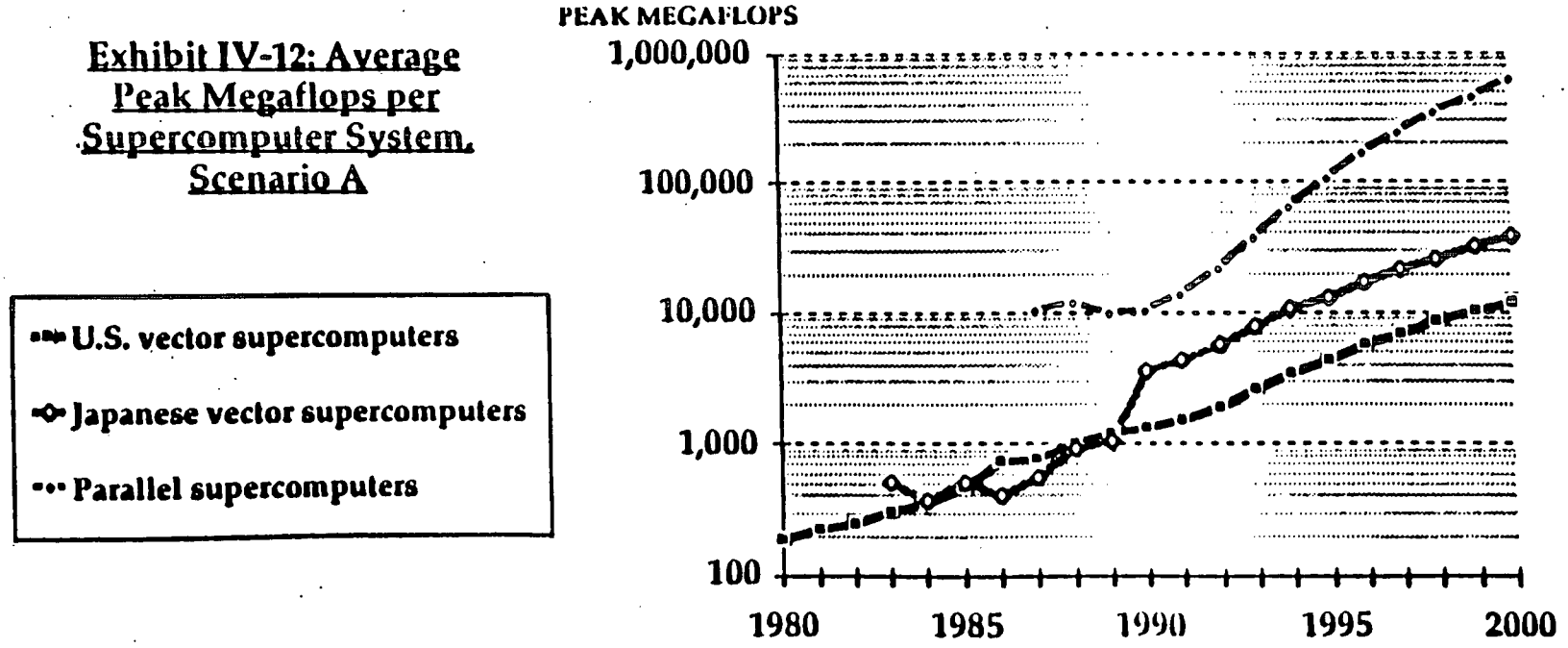
Exhibit IV-11: Price/Performance for Supercomputers, Scenario A



Despite the decrease in price per megaflops, average prices for supercomputer systems have actually increased a few percentage points per year historically. One reason for this is inflation. Another is that average system size has grown significantly, not only because of advances in component technology, but especially because of expanded use of multiprocessing -- that is, parallelism. We expect these trends to continue in vector supercomputers, although at a slowed rate of increase in processing power after 1995, due to limitations in circuit speed. For parallel systems, on the other hand, technological advances should lead to accelerated growth rates in processing power, especially in the latter part of the decade.

Assumption #5: We assume that the average processing power per vector supercomputer system will continue to grow, albeit at decreasing rates. For parallel supercomputers, we assume a significantly higher rate of processing power growth than in vector supercomputers (see Exhibit IV-12).

Exhibit IV-12: Average Peak Megaflops per Supercomputer System, Scenario A



Like other devices, supercomputer systems do wear out. More commonly, however, they are retired, because they have become obsolete. Newer systems are more reliable, easier to maintain, more powerful, and easier to use. Even desktop systems may approach the processing power of supercomputers two or three generations past, so they are preferable to old supercomputers (but not necessarily new ones).

Assumption #6: We assume that retirement rates for supercomputer systems of all types will follow historical patterns exhibited by U.S.-made vector supercomputers.

These assumptions are sufficient to generate a projection of supercomputer usage for the next ten years.

As shown in Exhibit IV-13, Japanese vector supercomputers will overtake their U.S. counterparts in terms of total installed peak processing power by 1992. The reason, in large part, is that the latest generation of Japanese supercomputers have rated peak speeds which are significantly higher than those of American vector systems (see Exhibit IV-12). This explanation is confirmed by Exhibit IV-14, which shows the U.S. vendors holding onto the lead in the number of installed vector supercomputer systems until the end of the decade.

Exhibit IV-13: Installed Peak Megaflops, Scenario A

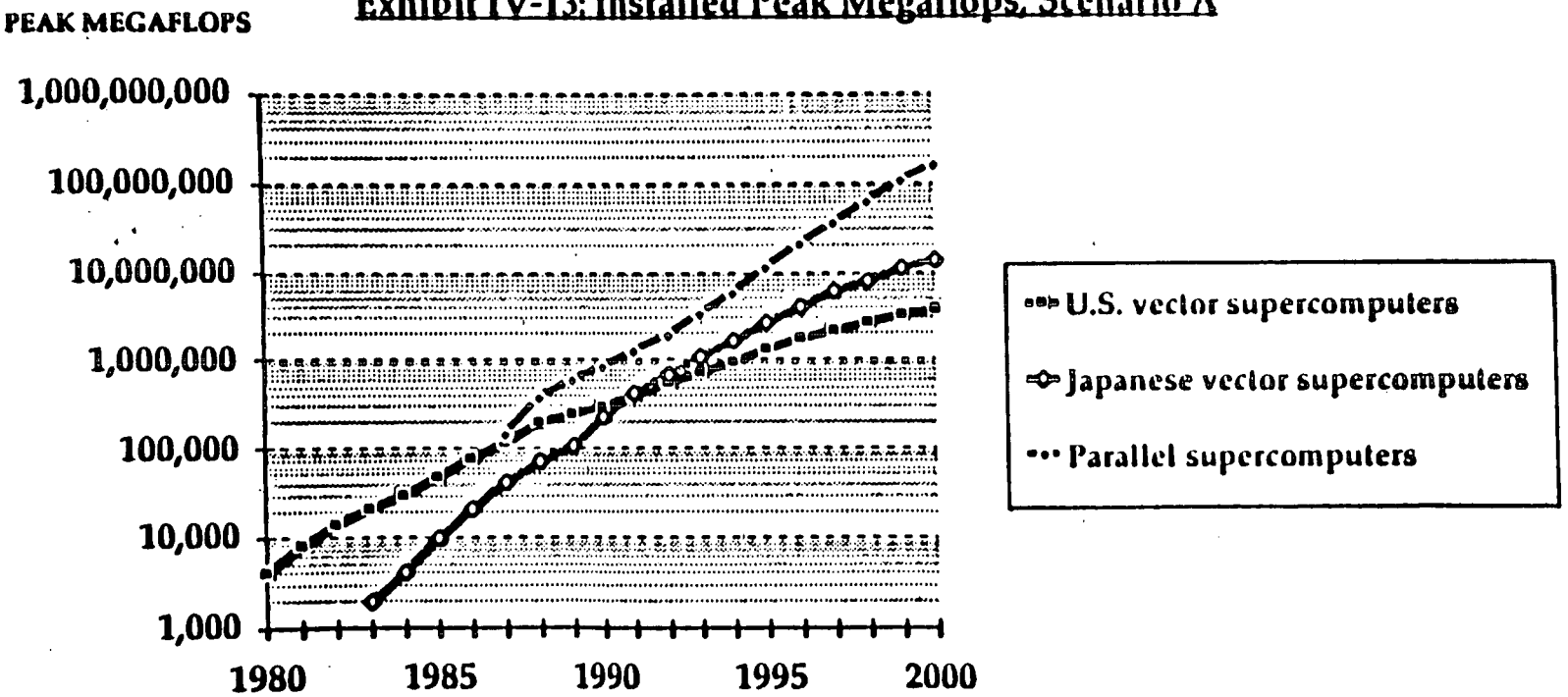
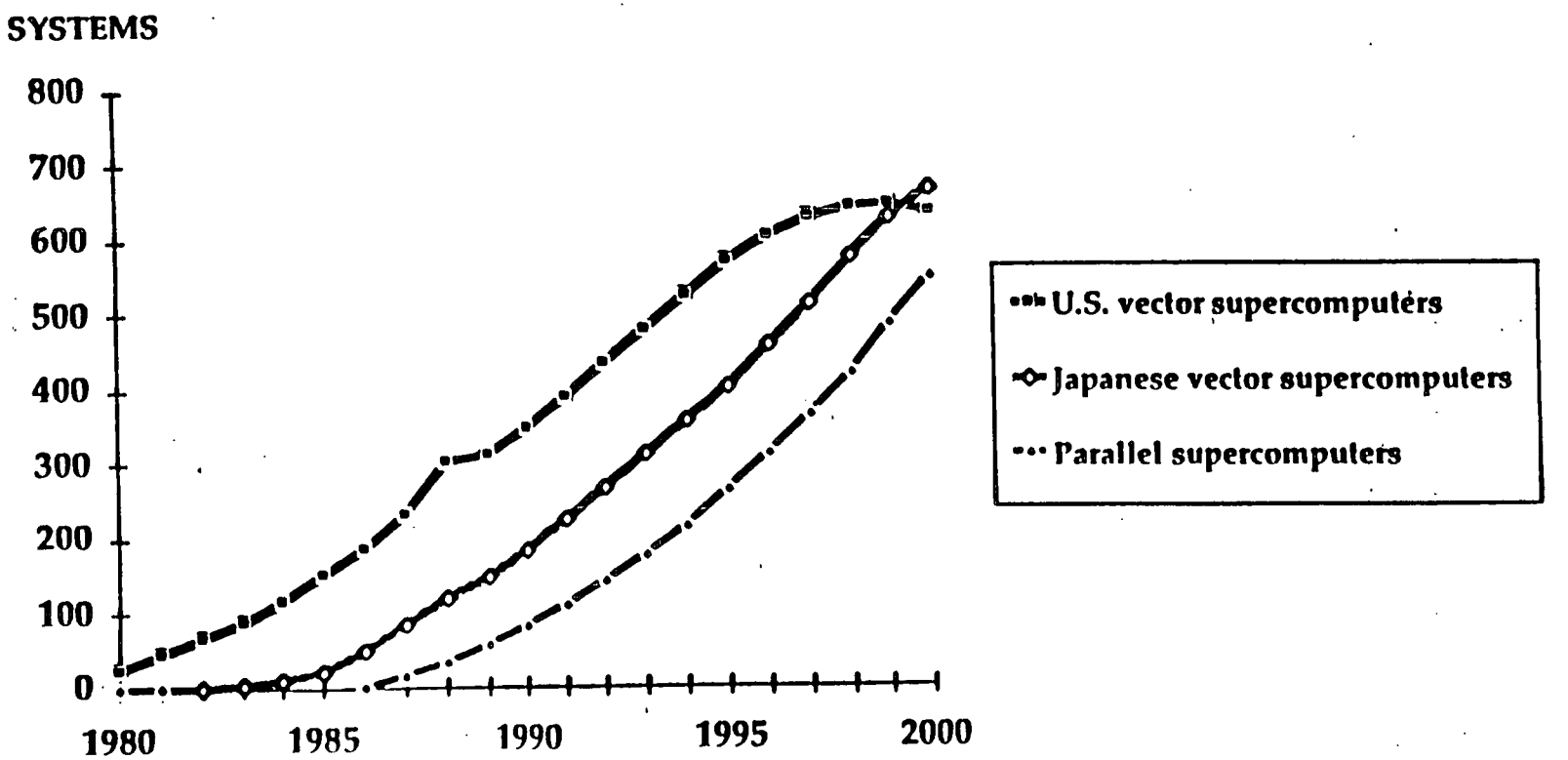
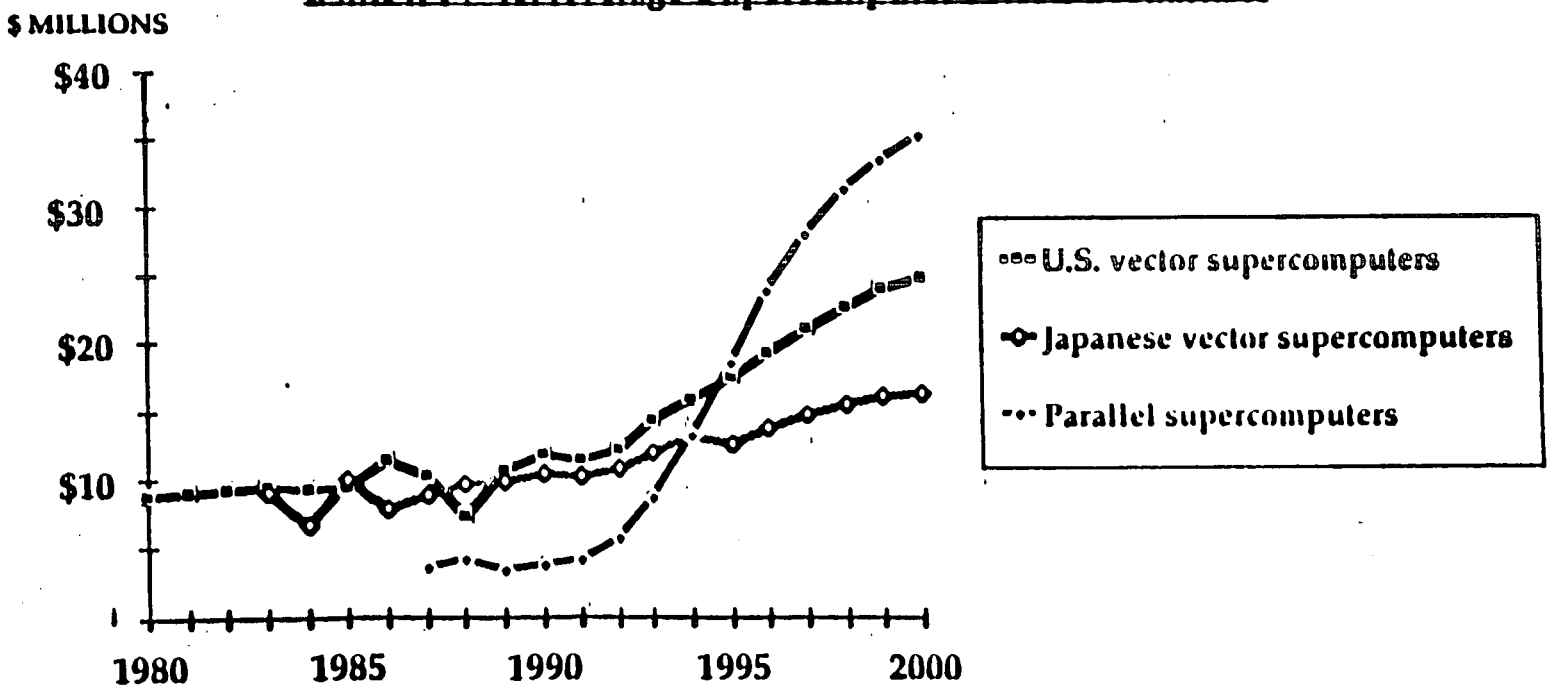


Exhibit IV-14: Installed Supercomputer Systems, Scenario A



As a kind of "reasonability check" on this analysis, we calculated the average prices for the three classes of supercomputer systems used in this scenario. As shown in Exhibit IV-15, the average price for a U.S.-made vector supercomputer will double by the year 2000, which is consistent with growing R&D costs and with expert projections that the price of the largest systems will increase from their present level of about $20 million to more than $40 million. (The actual price increase will be substantially less, because the pricing model used here has a built-in inflation factor of about 3 percent per year as a result of using "current dollars" in all price data going back to 1980.) The price of Japanese-made vector systems, however, will not increase as much, primarily because they are expected to lag behind U.S. systems in the degree of parallelism. The average price of parallel systems, on the other hand, will increase dramatically, as the largest systems approach the teraflops (1,000 gigaflops) level by the end of the decade.

Exhibit IV-15: Average Supercomputer Prices, Scenario A



Although this scenario clearly shows the flowering of parallel supercomputing, megaflops shipments for vector systems will begin to fall off after 1995 (see Exhibit IV-16). This decline in shipments, coupled with continuing improvements in price/ performance, will result in a downturn in revenue for vector supercomputer systems (see Exhibit IV-17). There is little chance of escaping this: the intense competition among the Fujitsu, Hitachi, and NEC will virtually insure that the price/performance of Japanese supercomputers will continue to fall as fast as the technology permits, and U.S. vendors will have no choice but to attempt to match the Japanese in worldwide markets.

Exhibit IV-16: Supercomputer Megaflops Shipments, Scenario A

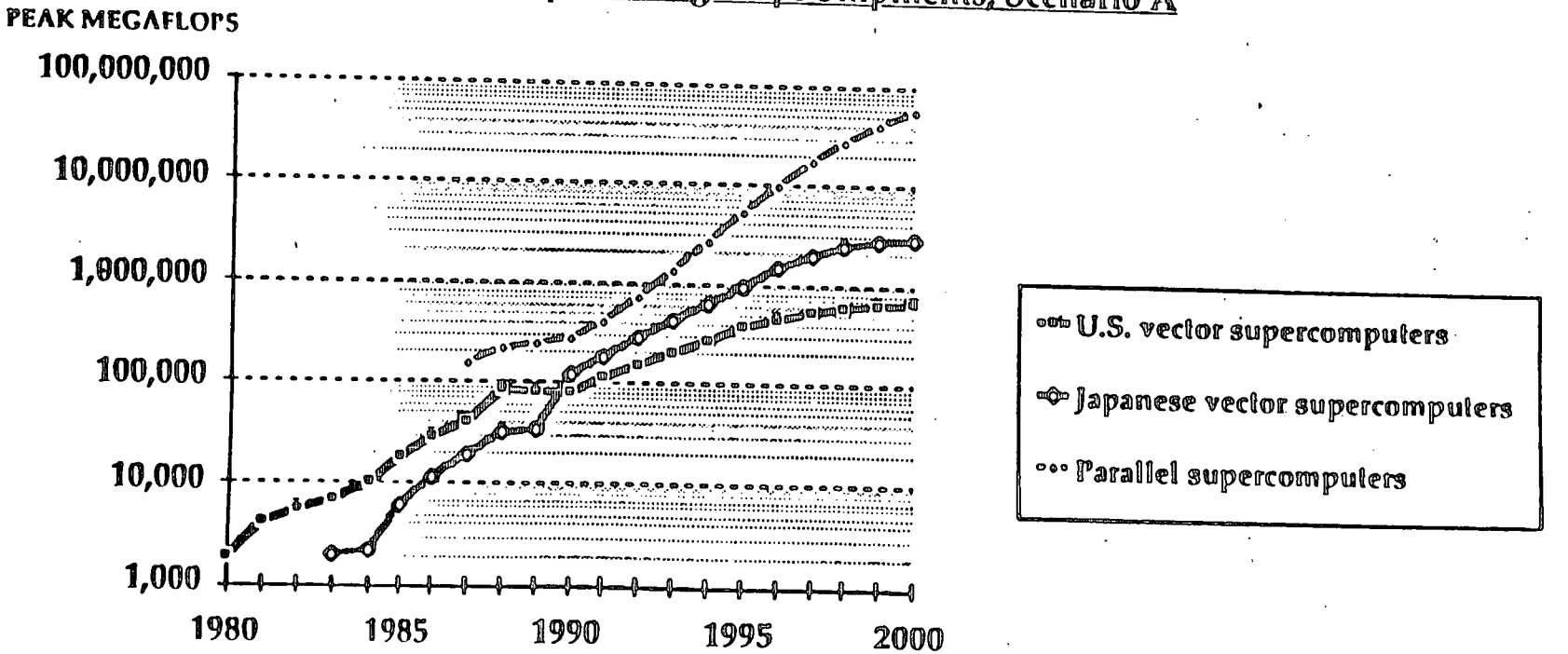
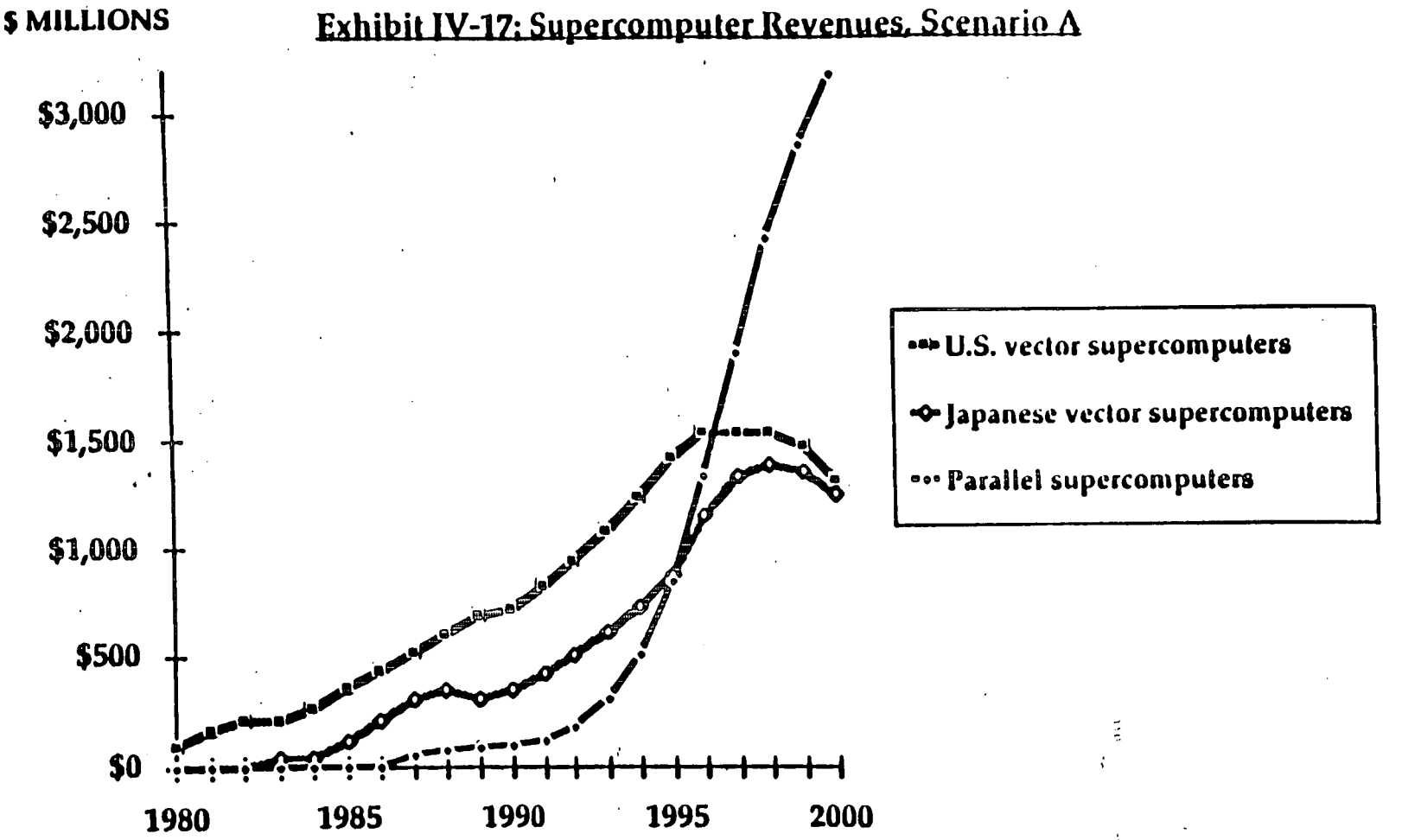


Exhibit IV-17: Supercomputer Revenues, Scenario A



As a result of this trend, we expect that companies which now focus on vector supercomputers will shift to parallel supercomputers in the latter 1990s -- that is, if they can afford the development costs. Based upon the size and resources of the Japanese companies (see Exhibit III-1), we have little doubt that they will be able to do this -- and judging from recent press announcements of an impending MITI-sponsored research project in massively parallel systems, they will also have the technology base and the impetus to do so. But we are concerned that the narrower base of U.S. vendors will not allow them to keep pace -- which lends an element of urgency to the Federal HPCC Program.

To complete Scenario A, we now turn to anticipated changes in usage patterns over the coming decade. As shown in Exhibits IV-18 and IV-19, we project increasing usage, but at declining rates of growth, which is typical of a maturing market. The academic decline is the most severe, reflecting the absence of strong Federal support, such as would be provided by the proposed HPCC Program. Industrial usage will provide the principal strength in the supercomputer market, but even that will be weakening in the later 1990s, unless an HPCC Program is implemented to develop new applications and to infuse computational science into the industrial R&D process at a more rapid rate than is now occurring.

Exhibit IV-18: Installed Supercomputer Systems, Scenario A

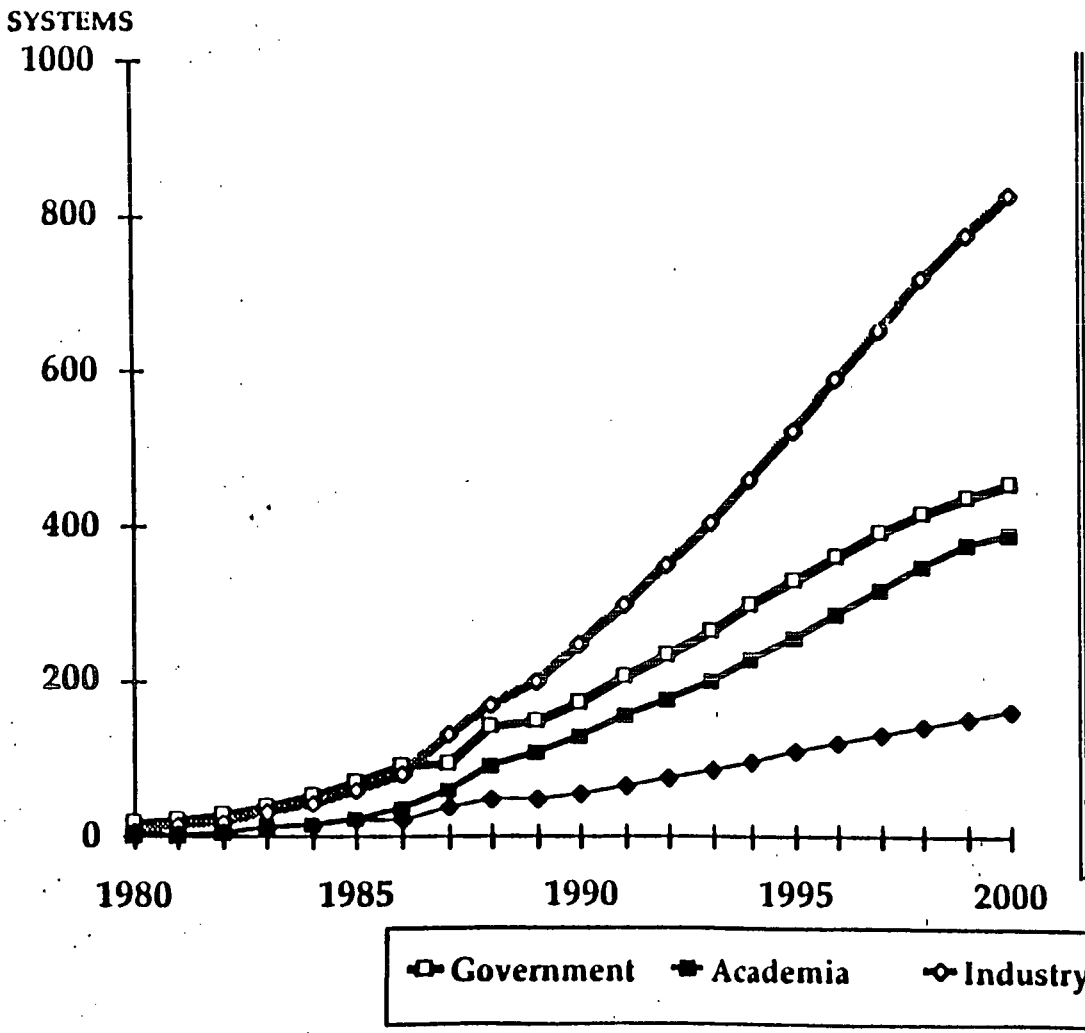
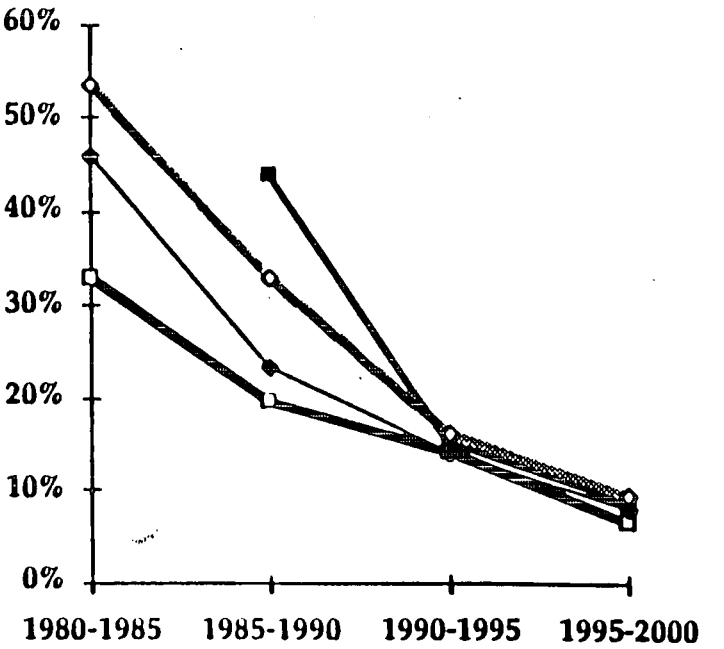


Exhibit IV-19: Growth Rates for Installed Supercomputer Systems, Scenario A



The final aspect of Scenario A is shown in Exhibits IV-20 and IV-21, which focus upon U.S., European, Japanese, and other nations' usage of supercomputers. Once again, the outlook is rather bleak for America, with Japan surpassing the U.S. in total number of installed supercomputer systems by 1997.

Exhibit IV-20: Installed Supercomputer Systems, Scenario A

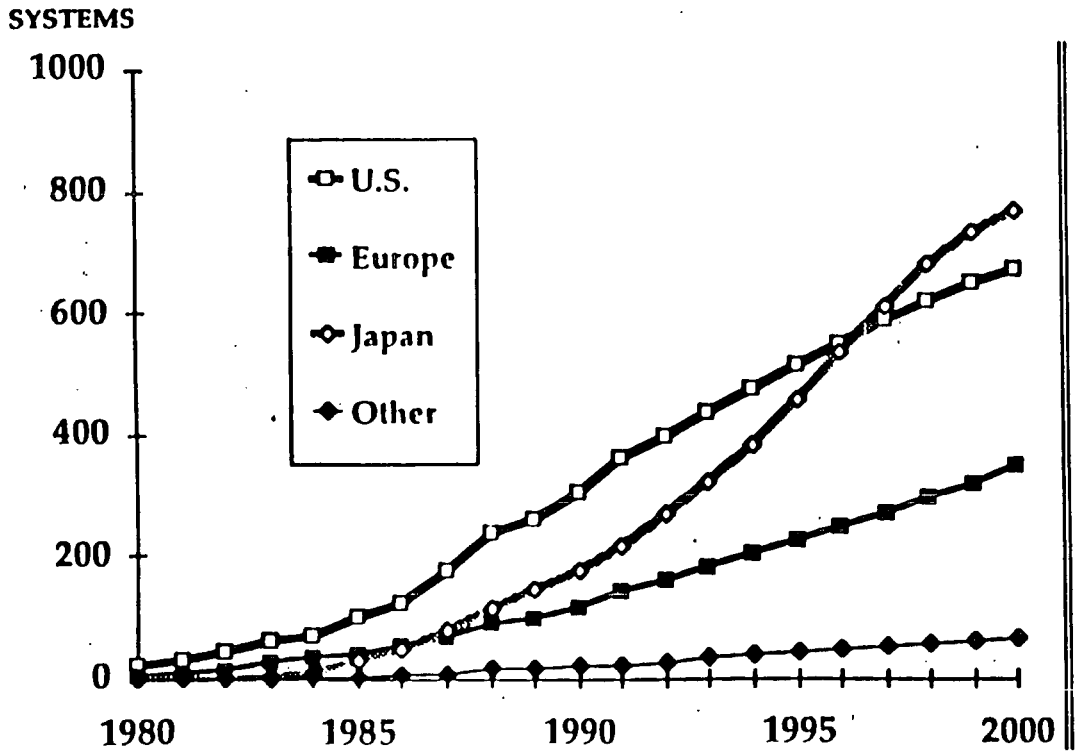
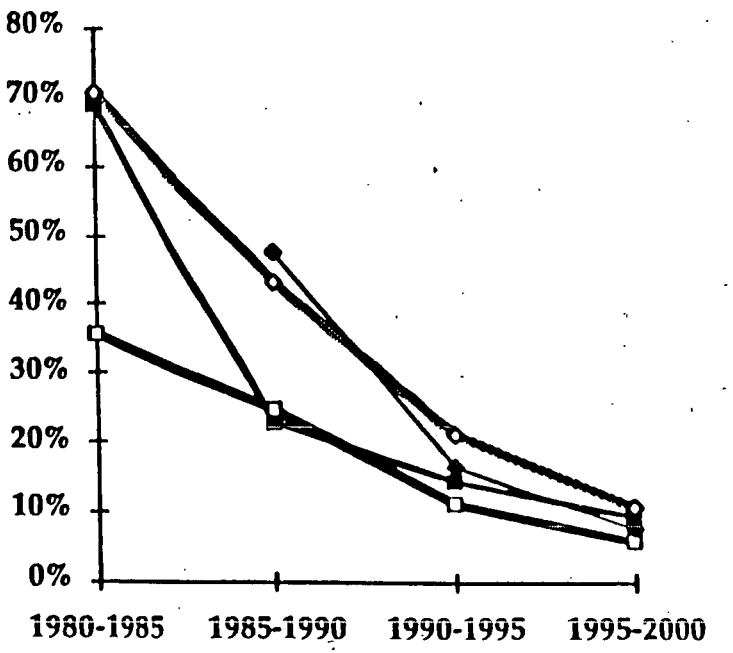


Exhibit IV-21: Growth Rates for Installed Supercomputer Systems, Scenario A



To summarize Scenario A, we project that:

- Installed supercomputer processing power (measured in peak megaflops) will increase more than 125-fold over the next decade (as compared with nearly triple that rate in the 1980s). Of the 176 million peak megaflops installed in the year 2000, 90 percent will be in parallel systems.
- The average system peak processing power will increase about tenfold for vector supercomputers -- to about 12 gigaflops for U.S.-made systems and about 38 gigaflops for Japanese-made systems -- and about sixty times for parallel systems -- to more than 600 gigaflops. The average price will be about $25 million for U.S.-made vector supercomputers, $16 million for Japanese-made vector supercomputers, and $35 million for parallel supercomputers.
- The average supercomputer price/performance will improve by a factor of 25 over the decade. A small part of this will come from price/performance improvements in the various types of supercomputer systems, but most of it will come from increasing usage of parallel systems, which have better price/performance than vector systems.
- Total processing power shipments (as measured in peak megaflops) will continue to grow throughout the decade, but growth rates for vector supercomputers will drop, causing revenues for vector systems to peak (at almost $3 billion) in 1998. Megaflops shipments for parallel systems will grow much faster than those for vector systems throughout the 1990s, and annual revenues for parallel systems will exceed those for vector systems by 1999.

... continued on next page

- The number of installed supercomputer systems will more than triple by the year 2000. Exhibit IV-22, shows how this installed base will be divided, as compared with today.

Exhibit IV-22: Installed Supercomputer Systems, Scenario A

Source	1990	2000
U.S. vector supercomputers	347 (57%)	640 (34%)
Japanese vector supercomputers	183 (30%)	669 (36%)
Parallel supercomputers	81 (13%)	552 (30%)
User		
Government	174 (28%)	463 (25%)
Academia	130 (21%)	402 (22%)
Industry	250 (41%)	833 (45%)
In-house	57 (9%)	163 (9%)
U.S.	301 (49%)	683 (37%)
Europe	115 (19%)	345 (19%)
Japan	174 (28%)	768 (41%)
Other	21 (3%)	65 (3%)
TOTAL	611	1,861

Of course, things don't have to turn out this way: for example, consider Scenario B.

SCENARIO B

As suggested above, the differences between Scenario A and Scenario B are in:

- Direction of HPC development and utilization and
- Rate of change in HPC development and utilization

as a result of the Federal HPCC Program. As in Scenario A, we focus upon supercomputers.

<u>Assumption #1:</u> Supercomputers are grouped the same as in Scenario A: • U.S. vector supercomputers, • Japanese vector supercomputers, and • Parallel supercomputers.

Assumptions #2 and #3: We assume that demand for supercomputer systems of both the vector and parallel varieties will be increased (above that in Scenario A) by the HPCC Program components concerned with "Evaluation of Early Systems" and "High Performance Computing Research Centers." All funding for "Early Evaluation" (amounting to $137 million over five years) will go toward the purchase of parallel supercomputers, while funding for "Research Centers" ($201 million over five years) will be used for (U.S.-made) vector supercomputers and parallel supercomputers, tending more to the latter over time. We also assume that Federal funding in these areas will precipitate increased state government expenditures as well, albeit at lower levels.

Although all of these systems would be installed in academic and government facilities (mostly the former), we also postulate increased industrial demand for supercomputer systems in Scenario B, resulting from erosion of the obstacles cited in Chapter III by the education and networking components of the HPCC Program as well as the "technology transfer" components of the Program. Here, the emphasis will be more on (U.S.-made) vector supercomputers in the near term, although parallel systems will also gain popularity in the industrial sector in the late 1990s, as a result of academic and government laboratory development efforts supported by the HPCC Program.

In this context, therefore, the net effect of the Federal HPCC Program would be to:

- **Stimulate (U.S.) demand for supercomputers; and**
- **Accelerate the rate of development of parallel supercomputers.**

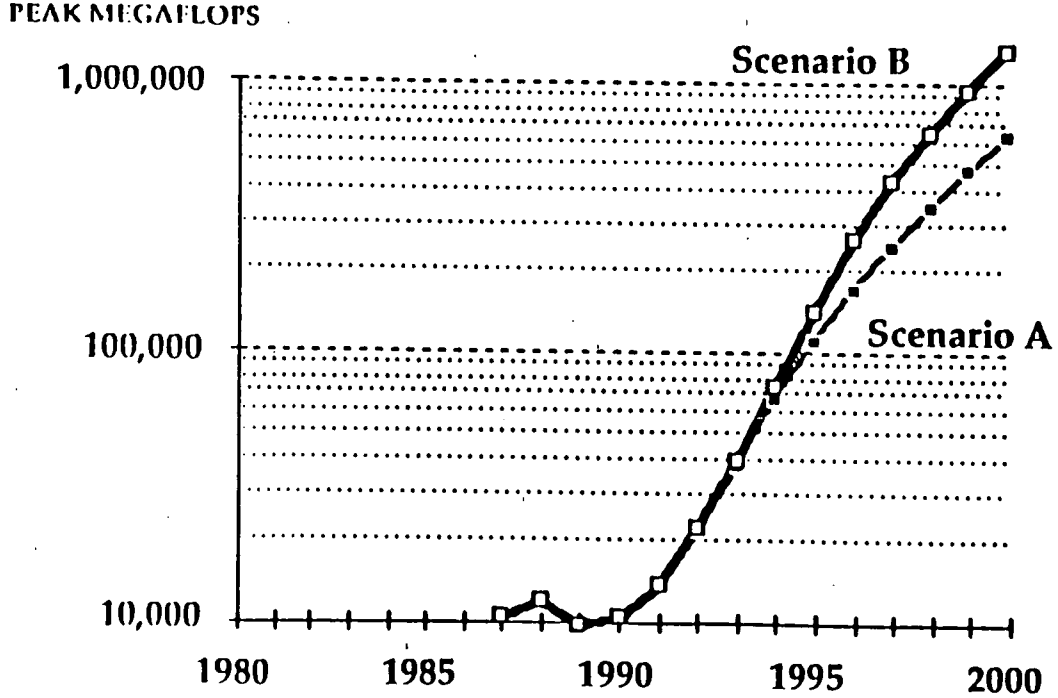
Combining these increments with the Scenario A "baseline" (given in Exhibit IV-22) yields the following Scenario B projections.

Exhibit IV-23: Installed Supercomputer Systems in the Year 2000, Scenarios A and B

	Scenario A	Scenario B
Source		
U.S. vector supercomputers	640 (34%)	754 (35%)
Japanese vector supercomputers	669 (36%)	669 (31%)
Parallel supercomputers	552 (30%)	750 (34%)
User		
Government	463 (25%)	488 (22%)
Academia	402 (22%)	518 (24%)
Industry	833 (45%)	984 (45%)
In-house	163 (9%)	183 (8%)
U.S.	683 (37%)	995 (46%)
Europe	345 (19%)	345 (16%)
Japan	768 (41%)	768 (35%)
Other	65 (3%)	65 (3%)
TOTAL	1,861	2,173

Assumption #5: We assume that the average peak processing power per vector supercomputer system will continue to grow at the same rates used in Scenario A. For parallel supercomputers, we assume that the peak processing power per system will grow more rapidly than in Scenario A, with the largest systems exceeding 1 teraflops by 1996 (see Exhibit IV-24).

Exhibit IV-24: Peak Megaflops per Parallel Supercomputer System, Scenarios A and B

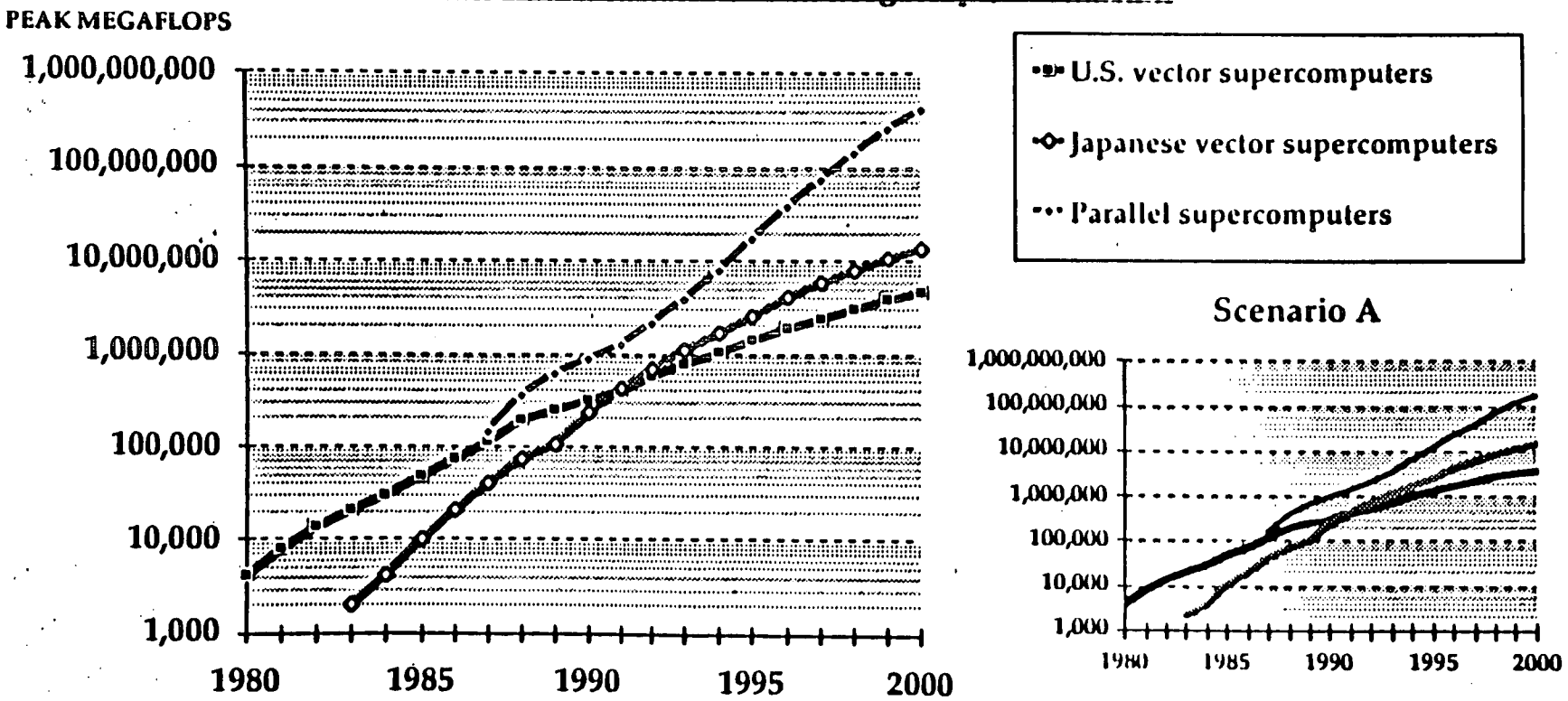


SCENARIO B

Assumption #6: We assume that retirement rates for supercomputer systems of all types will be the same as in Scenario A.

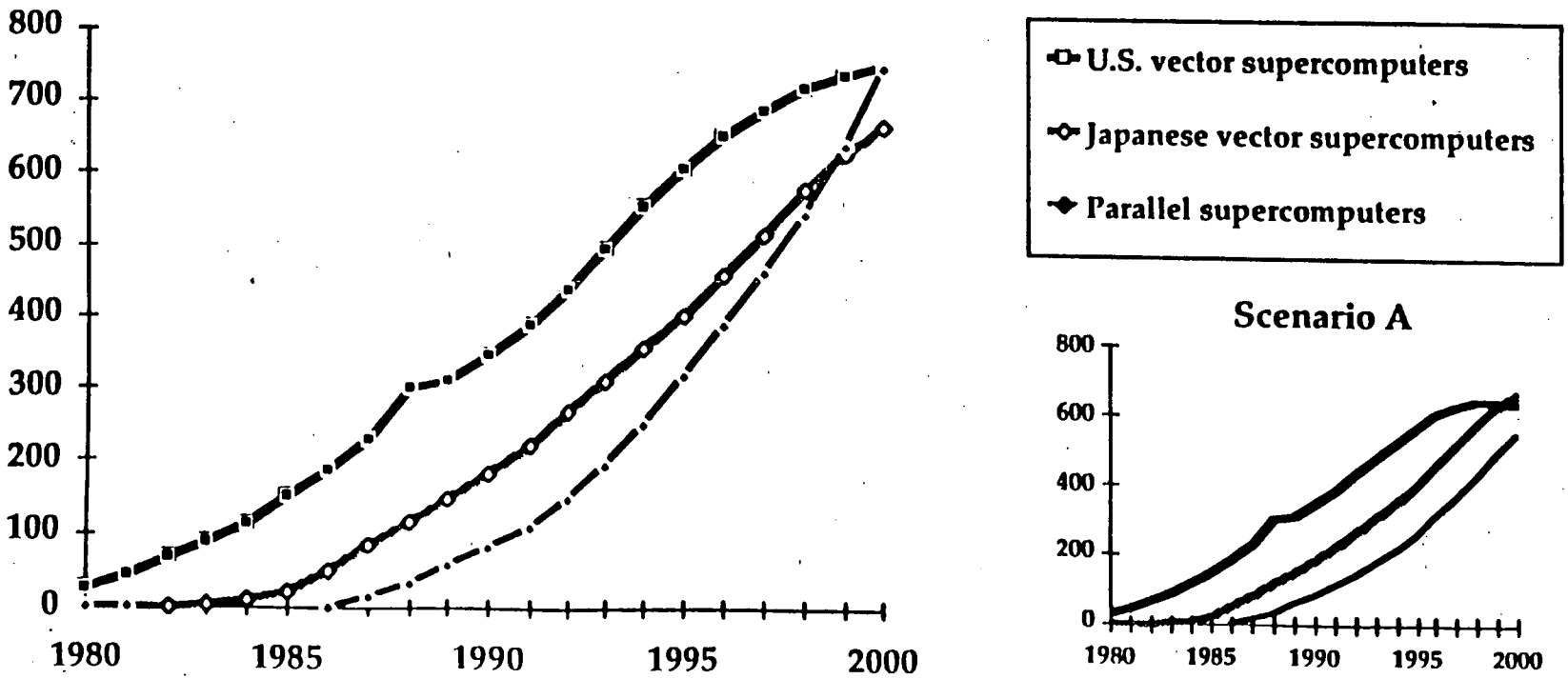
As for Scenario A, these assumptions are sufficient to generate a projection of supercomputer usage for the next 10 years.

Exhibit IV-25: Installed Peak Megaflops, Scenario B



The differences between Exhibit IV-25 and the corresponding exhibit (IV-13) for Scenario A appear rather slight, but this is because of the logarithmic scale used on the vertical axis. Actually, Scenario B shows a 250 percent gain in installed supercomputer power: 22 percent more peak megaflops in U.S. vector supercomputers and almost triple the peak megaflops in parallel supercomputers. As shown in Exhibit IV-26, the former number will be sufficient to keep U.S.-made vector supercomputers ahead of the Japanese competition in terms of installed systems -- assuming, of course, that the Japanese do not establish a counter-initiative to the U.S. HPCC program -- but what is more significant is the impact upon parallel systems usage: nearly 36 percent more systems of this kind will be installed by the year 2000 under Scenario B, as compared with Scenario A.

Exhibit IV-26: Installed Supercomputer Systems, Scenario B



SCENARIO B

The average price of (U.S.) vector supercomputer systems will be about 10 percent less under Scenario B (see Exhibit IV-27), thanks to the aforementioned economies of scale in production, but the average parallel supercomputer will cost as much as 30 percent more during the decade, decreasing to about 7 percent by the year 2000 (see Exhibit IV-28). (However, note that, as in Scenario A, these prices are in "current dollars," which assume about 3 percent per year inflation.) The reason for the latter is accelerated development of parallel systems technology under the HPCC Program, which will make possible the construction and -- more important -- the efficient utilization of significantly more powerful systems with larger-scale parallelism.

Exhibit IV-27: Average U.S. Vector Supercomputer Prices, Scenarios A and B

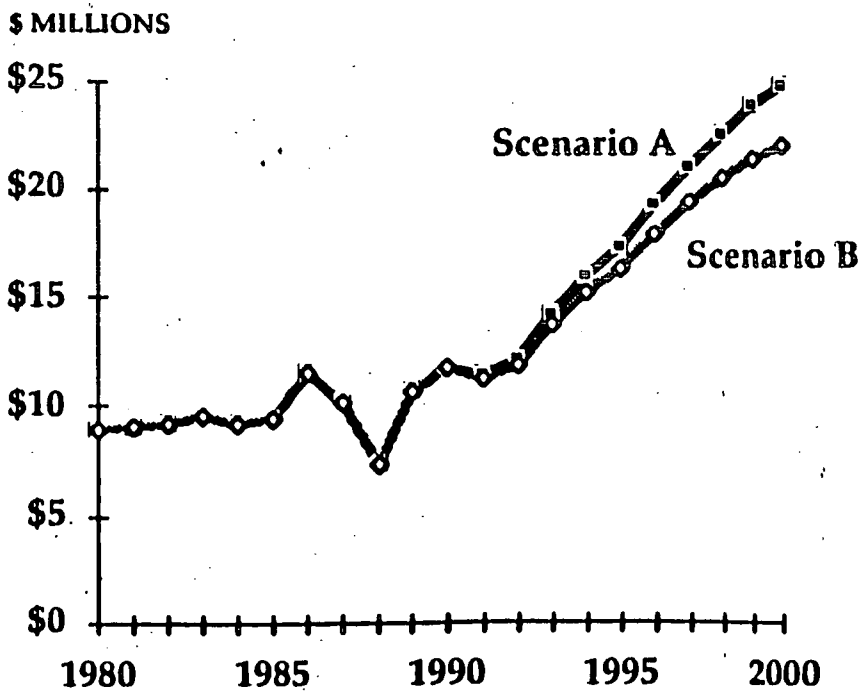
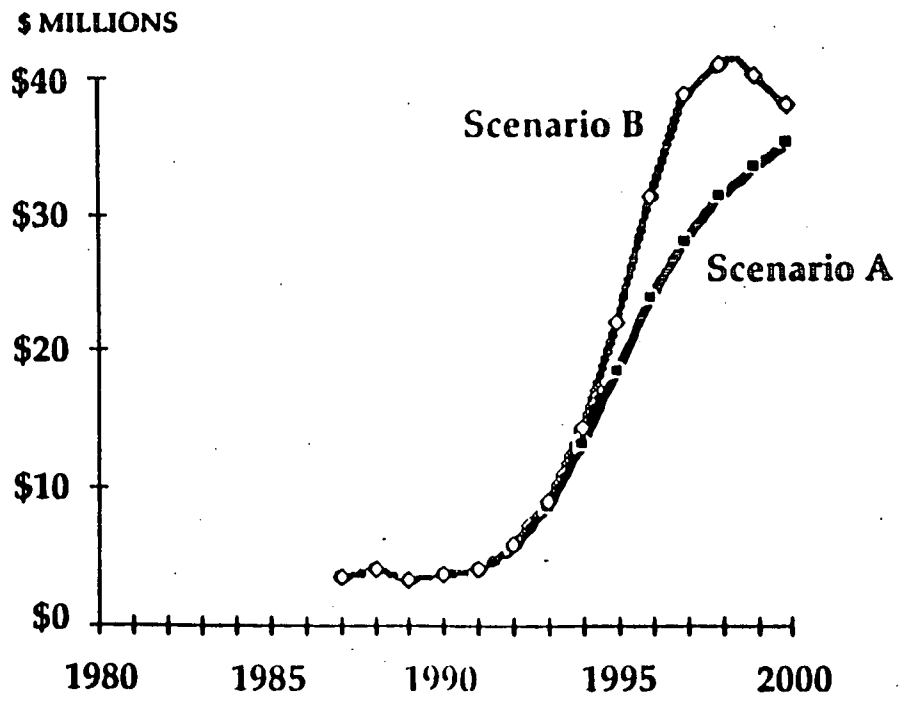


Exhibit IV-28: Average Parallel Supercomputer Prices, Scenarios A and B



The increase in U.S. supercomputer demand as a result of the HPCC Program will provide a 43 percent boost in U.S. vector supercomputer peak megaflops shipments in the year 2000, as compared with Scenario A, although signs of a maturing market will remain. For parallel systems, the increase in shipped megaflops will be even greater: a whopping 166 percent over Scenario A in the year 2000 (see Exhibit IV-29). These increases will translate into a 28 percent revenue improvement (over Scenario A) for vector systems and a 60 percent improvement in parallel systems revenue (see Exhibit IV-30).

Exhibit IV-29: Supercomputer Peak Megaflops Shipments, Scenario B

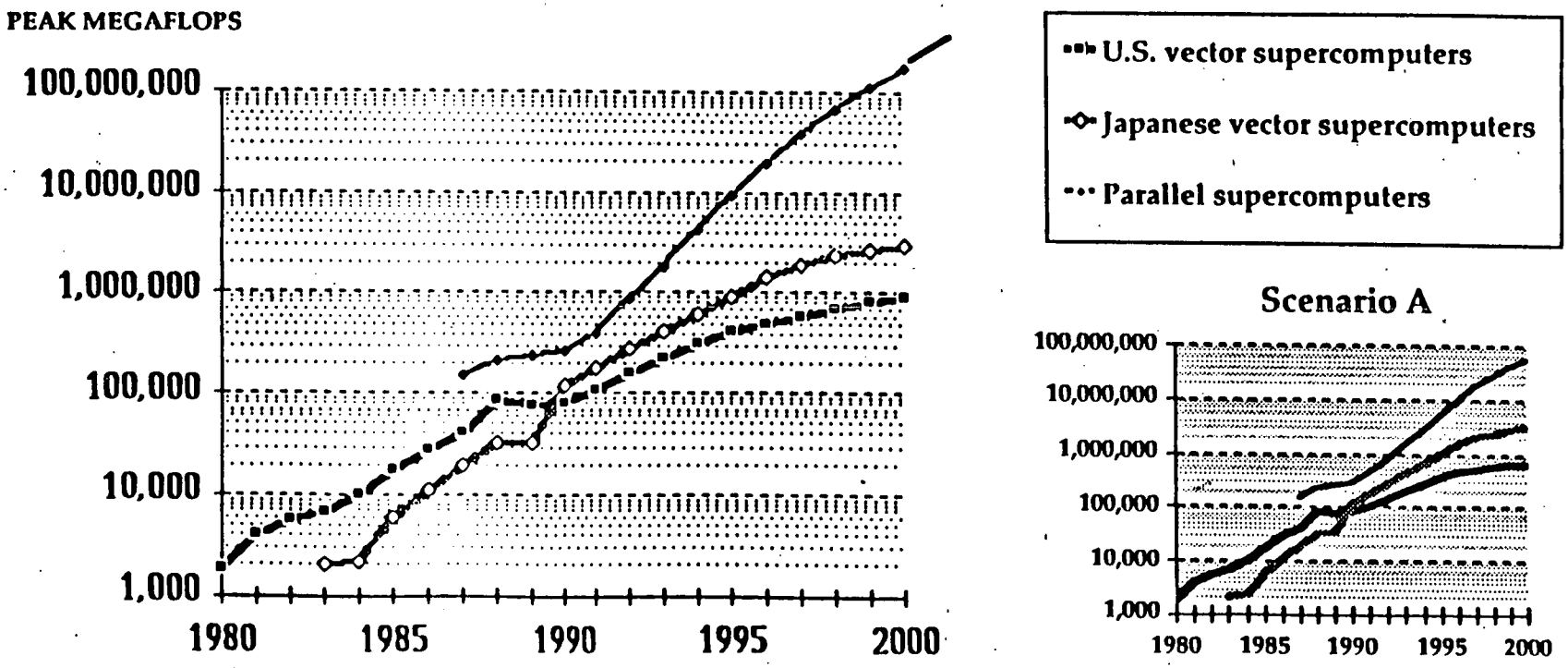
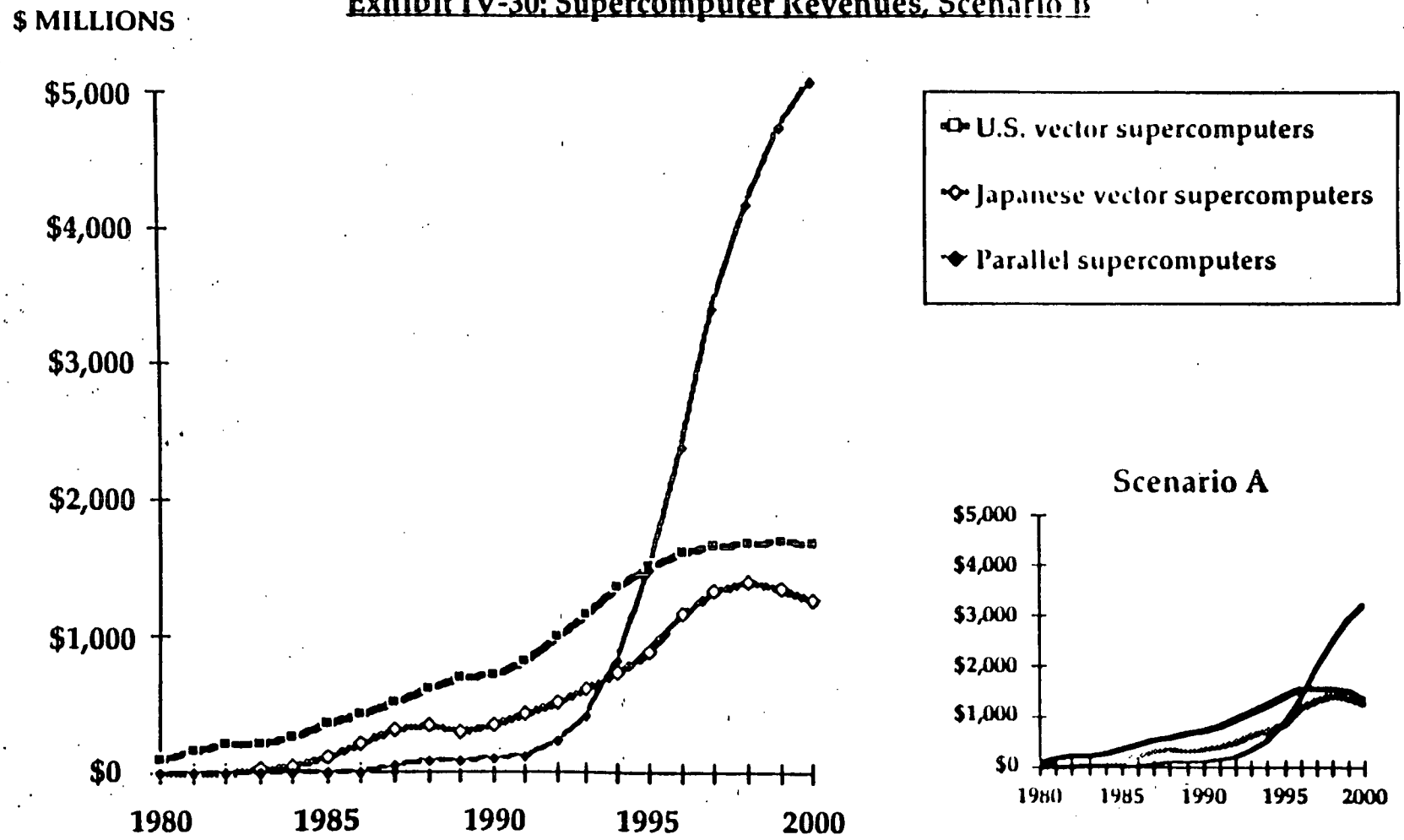
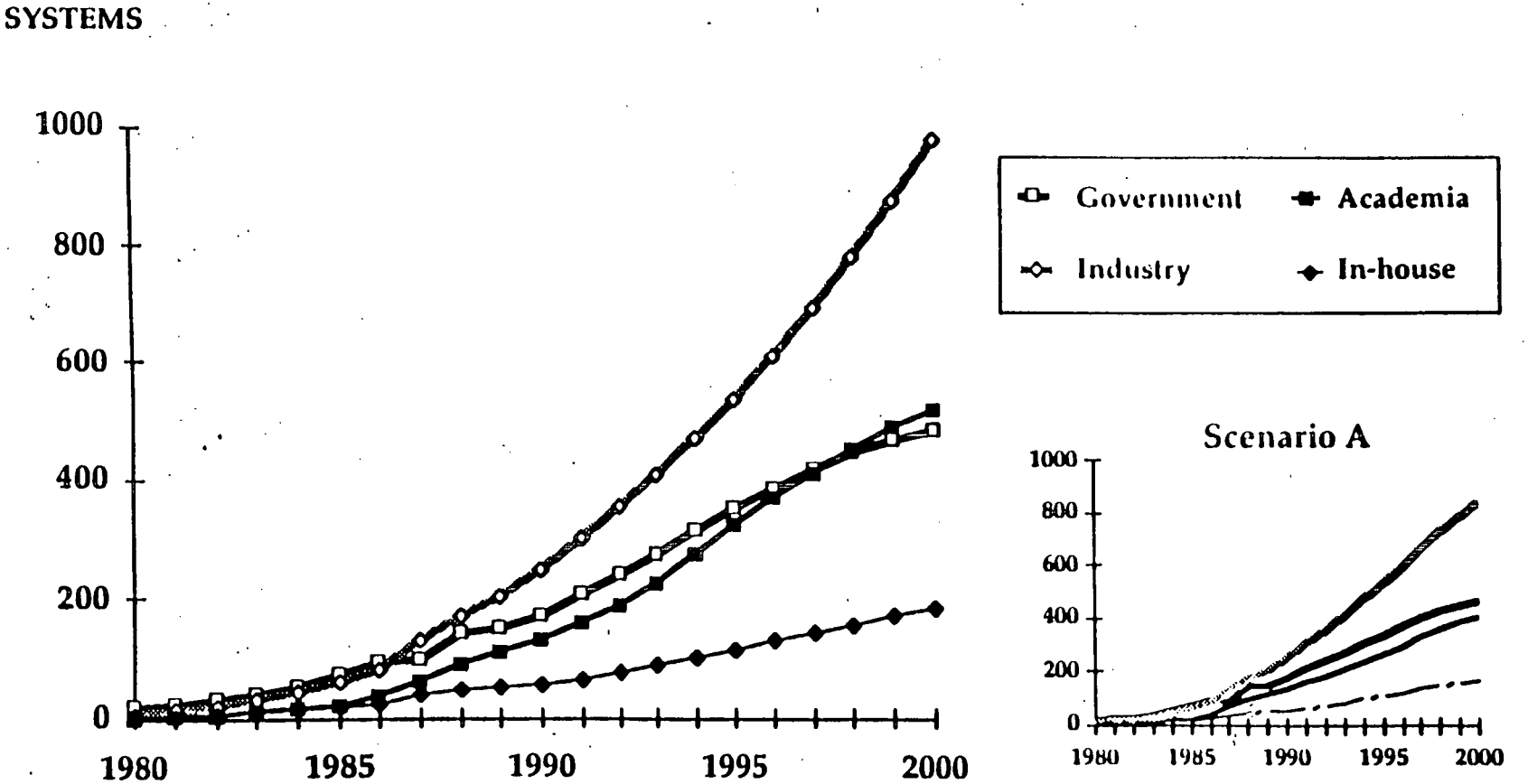


Exhibit IV-30: Supercomputer Revenues, Scenario B



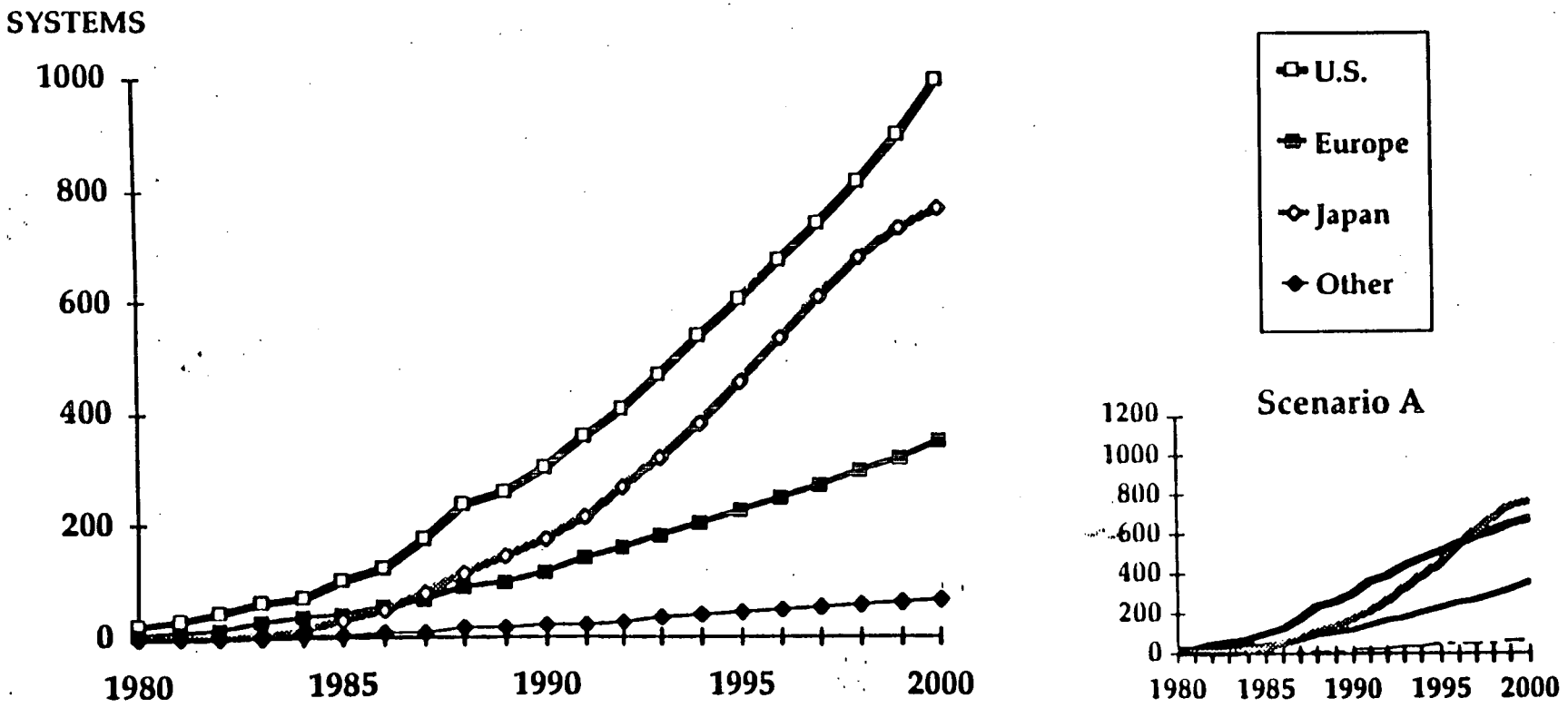
As shown in Exhibit IV-31, the HPCC Program will raise academic usage of supercomputers above that in the government by 1997: a gain of more than 30 percent above the Scenario A level. Government usage will also be expanded in Scenario B, but by only about 6 percent above Scenario A. Industrial usage will be 18 percent higher, and in-house usage about 12 percent more in Scenario B. This will be accomplished with very small changes in growth rates, ranging from about one percentage point higher for government to just over five percentage points higher for academia.

Exhibit IV-31: Installed Supercomputer Systems, Scenario B



However, as compared with Scenario A, the overall growth rate for U.S. supercomputer installations will be increased by 50 percent in Scenario B, which will result in the dramatic improvement in the U.S. position shown in Exhibit IV-32. U.S. leadership in supercomputer usage will be maintained, and this, in turn, will have a significant effect upon the economy as a whole, as will be explained in Chapter V.

Exhibit IV-32: Installed Supercomputer Systems, Scenario B



To summarize Scenario B, we project that:

- Installed supercomputer processing power (measured in peak megaflops) will be increased by a factor of over 300, to more than 440 million megaflops, by the year 2000. (Large as this may seem, the growth rate for the 1990s will be slightly less than that in the 1980s.) Of this, about 96 percent will be in parallel systems.
- In the average system, peak processing power will increase about tenfold for vector supercomputers -- to about 12 gigaflops for U.S.-made vector systems and about 38 gigaflops for Japanese-made vector systems -- and nearly 125-fold for parallel systems -- to more than 1 teraflops. U.S.-made vector supercomputers will cost an average of $22 million, Japanese-made vector supercomputers about $16 million, and parallel supercomputers slightly less than $38 million.
- The average supercomputer price/performance will improve by a factor of 55 over the decade.
- Total processing power shipments (as measured in peak megaflops) will continue to grow throughout the decade, but parallel supercomputers will begin to displace vector supercomputers by 1995. Annual worldwide revenues for the latter will top out at just over $3 billion in 1999, while revenues for parallel systems will be $5 billion by 2000.
- The number of supercomputer systems installed worldwide will approach 2,200 by 2000, with private industry (including in-house systems used by supercomputer vendors) accounting for more than half. The U.S. will lead the world in supercomputer usage with almost 46 percent of all installed systems.

HPCC PROGRAM IMPACT

A comparison of Scenarios A and B shows that the Federal HPCC Program will result in significant differences for the supercomputer industry in the year 2000:

- **39% greater supercomputer revenues;**
- **Almost 3 times more megaflops shipped;**
- **Almost 2.5 times more megaflops installed; and**
- **17% more systems installed.**

Perhaps most significant of all is the projected impact of the Federal HPCC Program on industry revenues (see Exhibit IV-33):

Under Scenario B, the cumulative supercomputer industry revenues for the 1990-2000 period will be

$10.4 billion

greater than under Scenario A.

This represents a 28 percent increase over Scenario A. It is more than five times greater than the projected $1.917 billion expenditure for the HPCC Program.

In addition, there would be increased revenues in other segments of the High Performance Computing industry as well: minisupercomputers, workstations, networking systems, and software. We have not calculated these increases, but we are confident that their total will be considerably greater than the impact upon the supercomputer industry. (At present, worldwide minisupercomputer revenues are about half those for supercomputers, as are revenues for mainframes with auxiliary vector processors. Revenues for high performance workstations are about $4 billion.)

Exhibit IV-33: Worldwide Supercomputer Revenues, Scenarios A and B

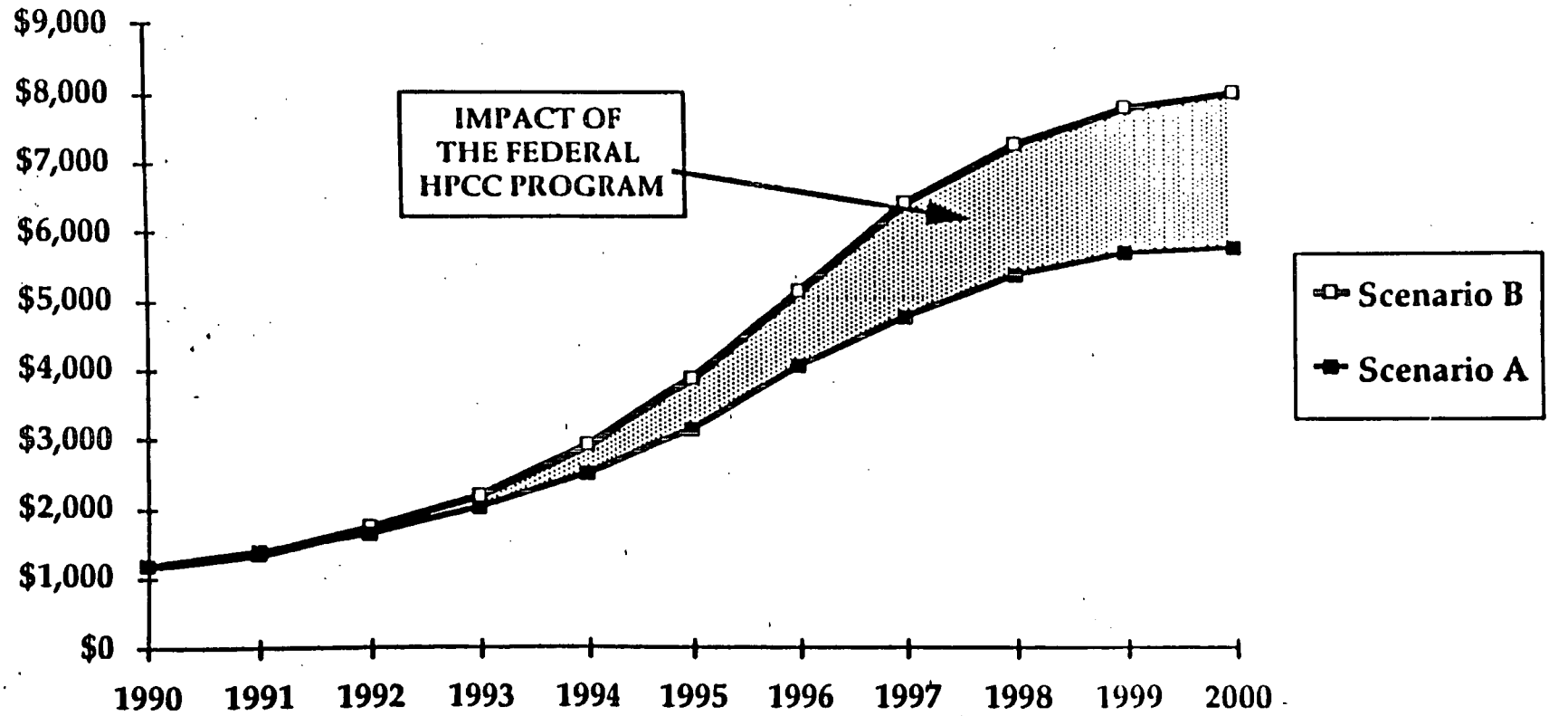


Exhibit IV-34: Worldwide Installed Peak Megaflops, Scenarios A and B

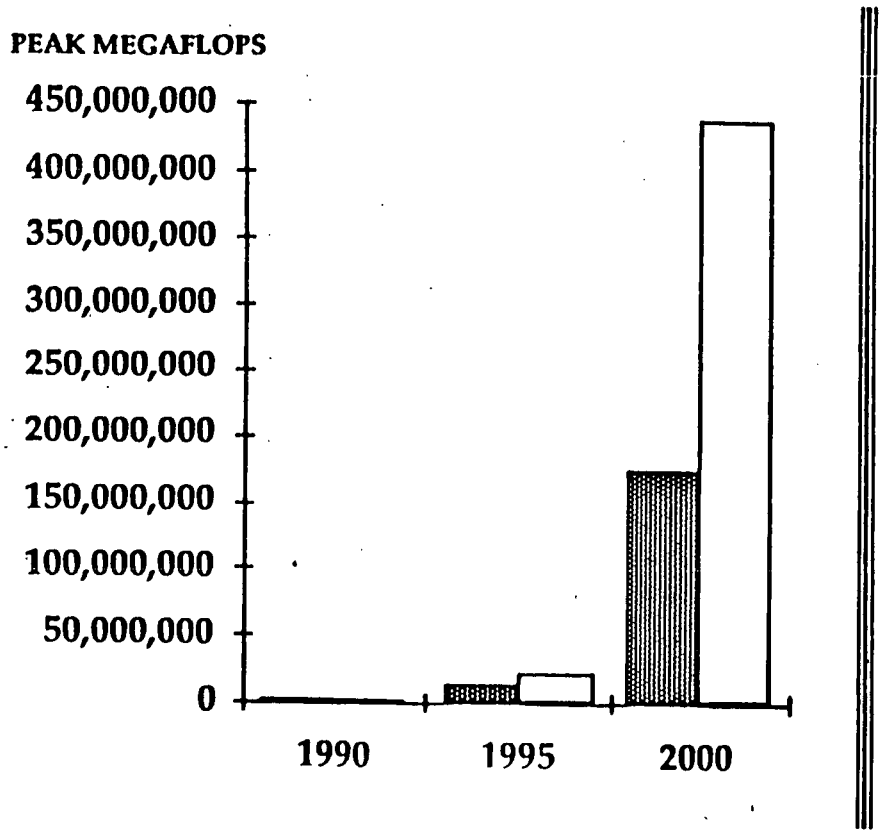
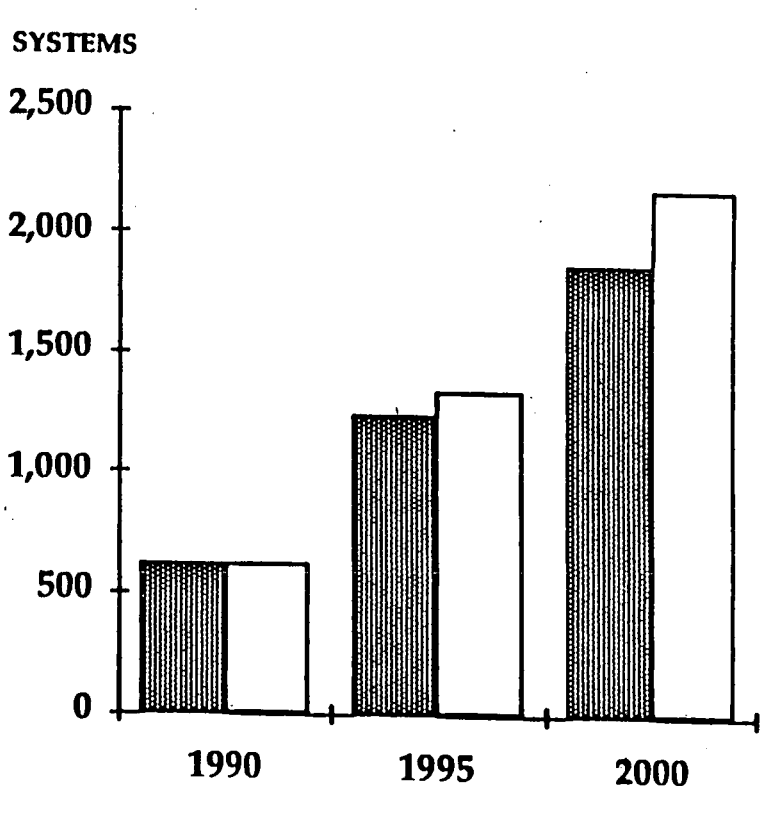


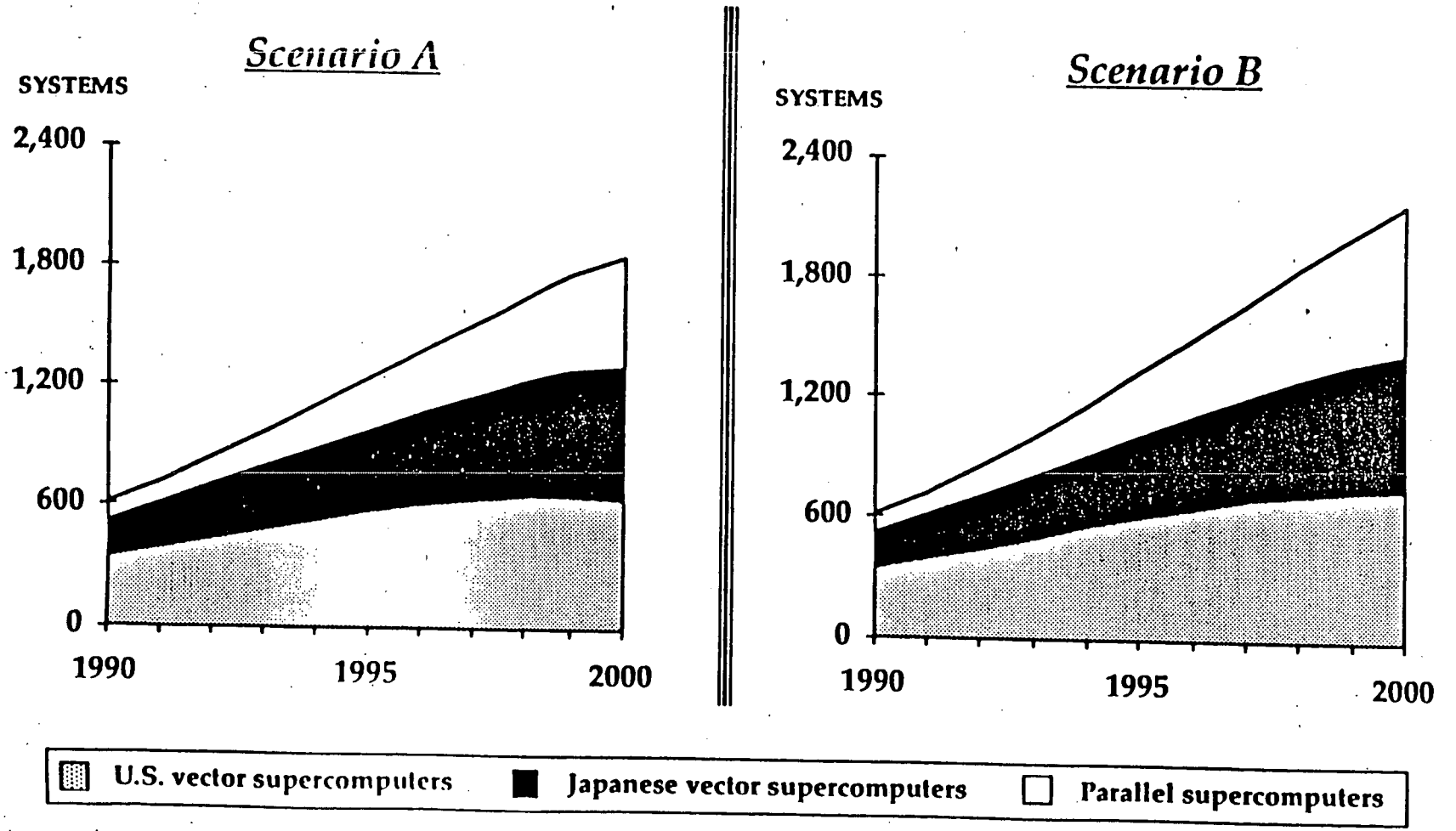
Exhibit IV-35: Worldwide Installed Supercomputer Systems, Scenarios A and B



As explained earlier, we expect that academia and (to a lesser extent) government will be the direct beneficiaries of HPCC Program funds for prototype evaluation and HPC research centers. In particular, we project that the HPCC Program will increase the number of supercomputer systems in academic use by the year 2000 by nearly 30 percent and the number in government use by more than 5 percent over what would otherwise be installed. Industrial usage (excluding in-house systems used by supercomputer vendors for their own R&D, marketing, etc.) will be boosted by about 18 percent, and the total number of supercomputer systems installed in the United States will be increased by 46 percent. As shown in Exhibit IV-32, this should be sufficient to keep the U.S. ahead of Japan in supercomputer installations, barring a major Japanese initiative to expand their usage as well. Looking at the situation another way, if the HPCC Program is not funded, Japanese supercomputer usage will surpass that in the U.S. by 1997 (see Exhibit IV-20).

In terms of the kinds of supercomputers deployed, Scenario B shows an increased usage of parallel systems, in line with our initial assumptions, although in both Scenarios we project that parallel systems revenues will exceed those of vector systems around 1998. But in terms of the parallel systems component of installed systems (see Exhibit IV-36), there is a larger difference between Scenarios A and B: almost 17 percent more systems will be installed in the year 2000 under Scenario B as compared with Scenario A. And we would expect that the differences between these two scenarios will sharpen even further after the turn of the century, because the results of the HPCC Program will continue to "trickle-down" for a long, long time.

Exhibit IV-36: Installed Supercomputer Systems by Type, Scenarios A and B



The distribution of computer systems among user groups -- government, academia, industry, and in-house -- varies only a small amount between Scenarios A and B. The principal difference is increased academic usage in Scenario B as a result of increased funding for HPC Research Centers under the HPCC Program (see Exhibit IV-37). (Of course, government and industrial usage would increase, too, but not as much.)

Exhibit IV-37: Installed Supercomputer Systems, by User

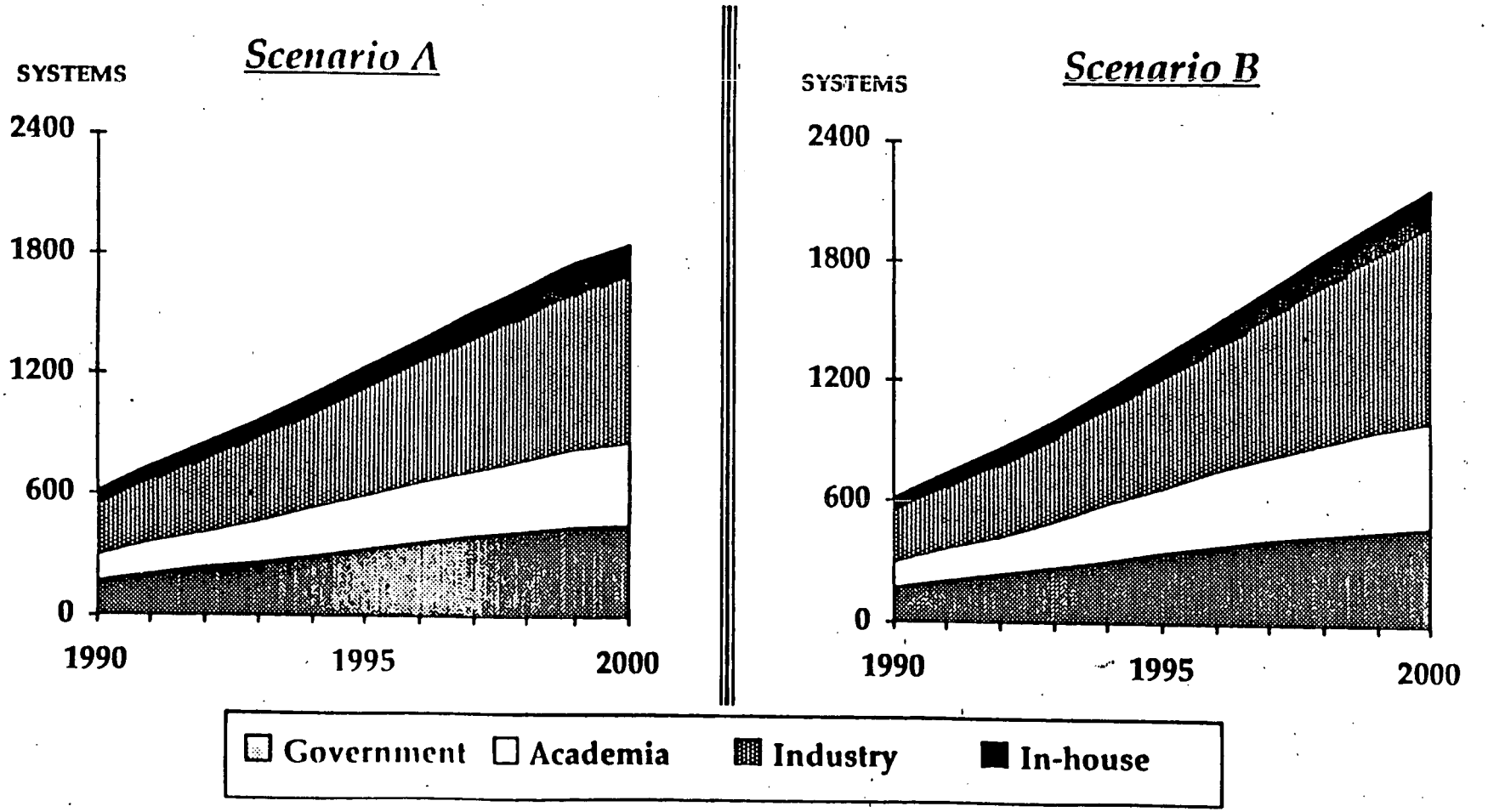
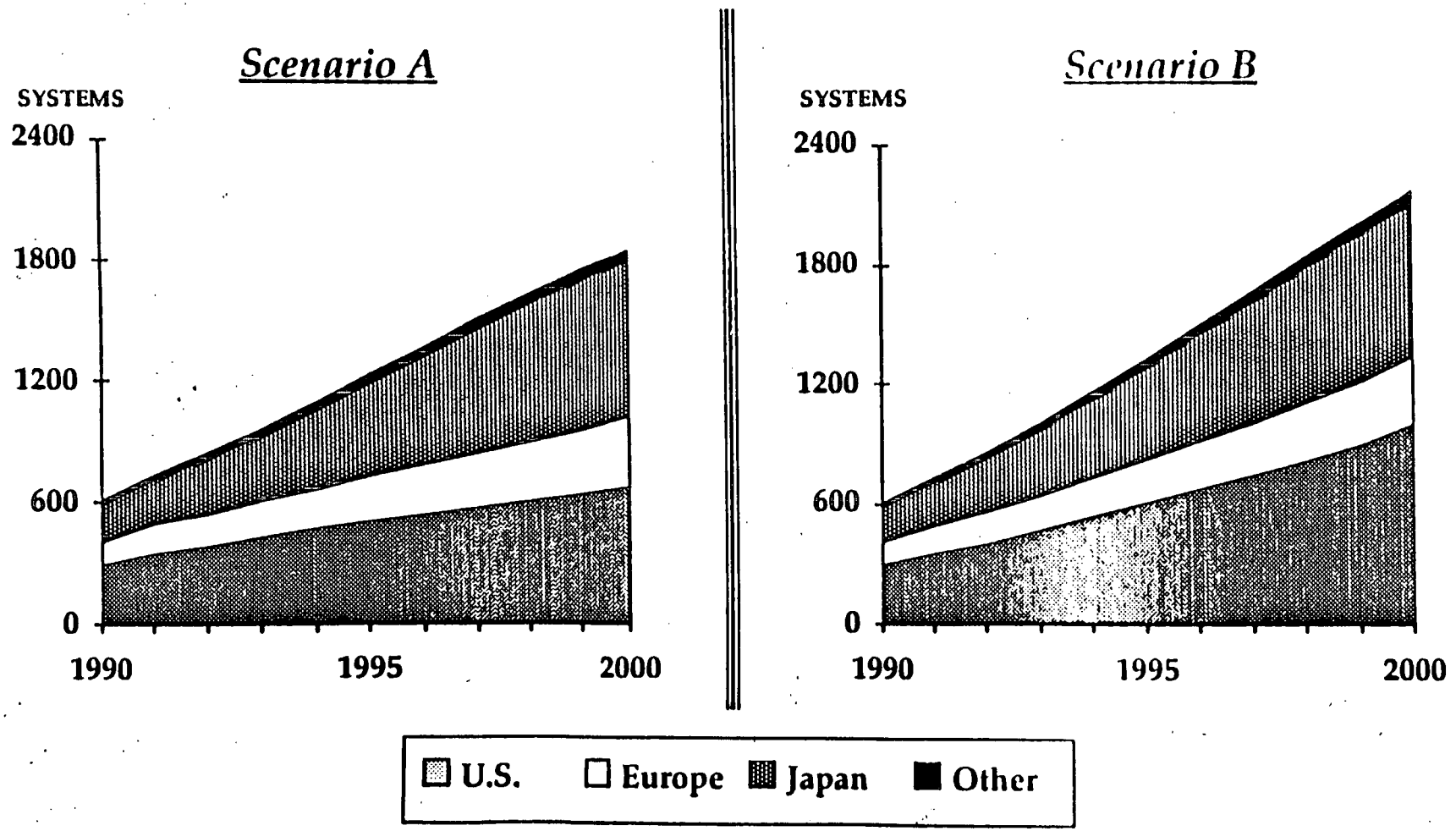


Exhibit IV-38: Installed Supercomputer Systems, by Country



We have not posited alternative scenarios for Japanese and/or European programs in HPC, because we regard Japan to be in "Scenario B mode" already, and it is much too soon to judge whether or not the recently-proposed EEC program in HPC will actually be established*. However, some additional foreign response -- "Scenario C?" -- to a U.S. HPC initiative is likely, because the history of French, German, and Japanese government investment in R&D has been strong and stable. But if that were to happen, a U.S. counter-response, especially in the form of an even greater increase in HPCC funding than that proposed by OSTP, would probably be inappropriate because of limitations on our own ability to absorb additional funds and to deploy additional resources effectively.

On the other hand, we generally concur with the HPCC Program, as articulated in the OSTP proposal of September 8, 1989, as the minimal reasonable government response to forestall the undesirable eventualities we foresee in Scenario A. We have not considered any other alternatives, such as partial HPCC funding, because we believe that there are "critical mass" and "synergy" principles which apply here. Therefore, any partial funding scenarios are likely to be equivalent in effect to Scenario A: that is, no funding at all.

*Cf. Report of the EEC Working Group on High-Performance Computing; Commission of the European Communities, 1991.

[This page has been left blank intentionally.]

CHAPTER V - HPC APPLICATIONS

V - HPC APPLICATIONS

This chapter provides two scenarios depicting the next ten years in High Performance Computing. It is arranged in two sections, as follows:

- **Productivity** - explains the expected penetration of supercomputing into major industry groups and describes the projected impact of the Federal HPCC Program on the United States economy over the next ten years.

- **Technological Leadership** - describes the degree of HPC sophistication in various technological application areas, at present and in the year 2000.

PRODUCTIVITY

The projected $1.9 billion cost of the Federal HPCC Program would probably be justified by the Program's projected impact on just the information industry alone: at least $10.4 billion in increased supercomputer industry revenues, plus corresponding revenue growth in other HPC segments and inestimable gains from increased vitality throughout the U.S. information industry as a result of "trickle-down." But even if this impact were nil, the Program would be more than justified by its likely benefits to users of HPC:

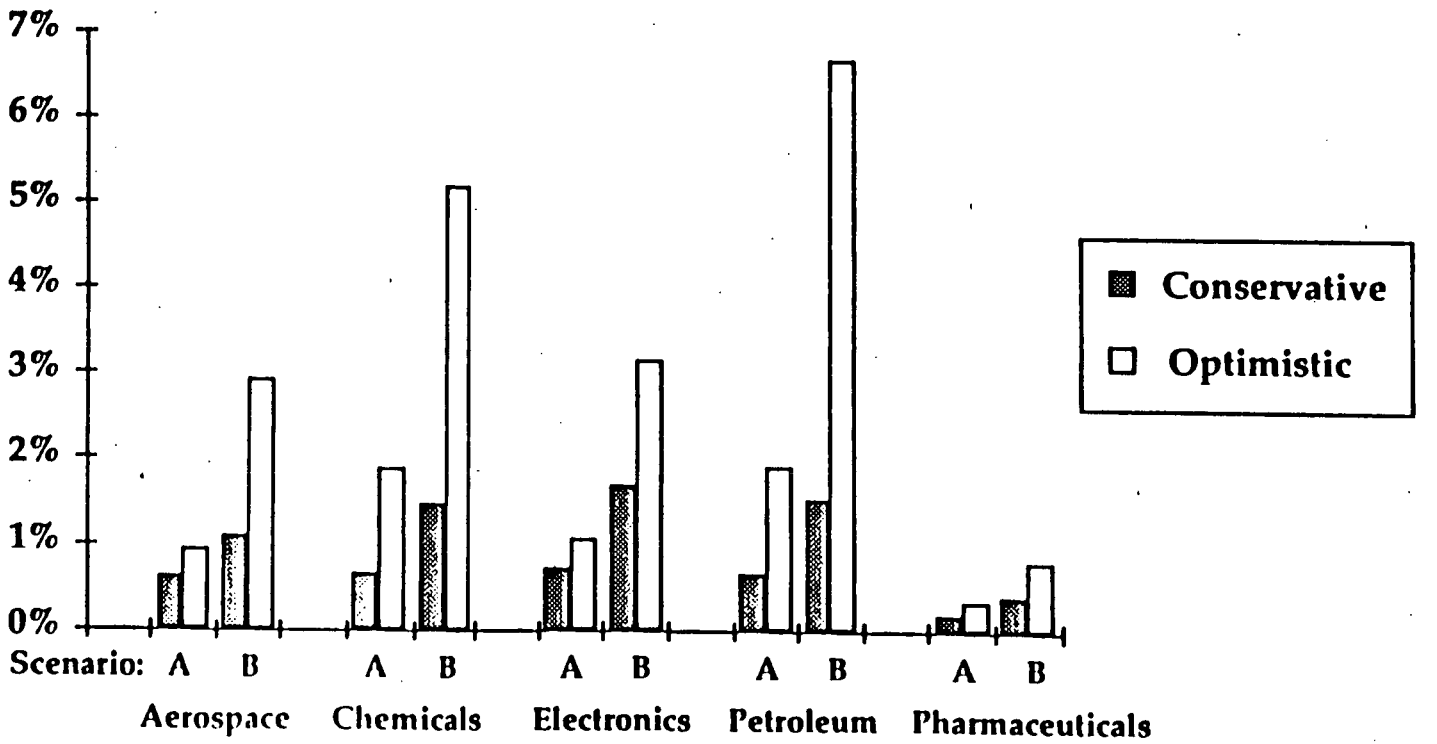
- **Significantly improved ability of industry to bring quality products and services to market quickly; and**
- **Greatly enhanced ability of U.S. scientists and engineers to meet the "Grand Challenges" and realize the other applications opportunities described in Chapter III.**

The first of these can be characterized as **improved R&D productivity** (which ultimately means improved overall productivity), and the second as **enhanced technological leadership.**

Our interviews of R&D managers in various industries which use HPC indicate that major gains in R&D productivity can be realized if the supercomputing capabilities which are projected for Scenario B in the preceding chapter do indeed come to fruition. Of course, these capabilities must be accompanied by commensurate improvements in other harder-to-quantify aspects of HPC usage -- software, networks, and (especially) trained people (see Appendix H) -- for this realization to occur, but there is unanimity that there will be significant benefits for HPC users. Naturally, the timing and projected levels of productivity gains vary considerably across different industries, but there is a surprising similarity in the expected gains at different companies -- at least, those we interviewed -- within the same industry.

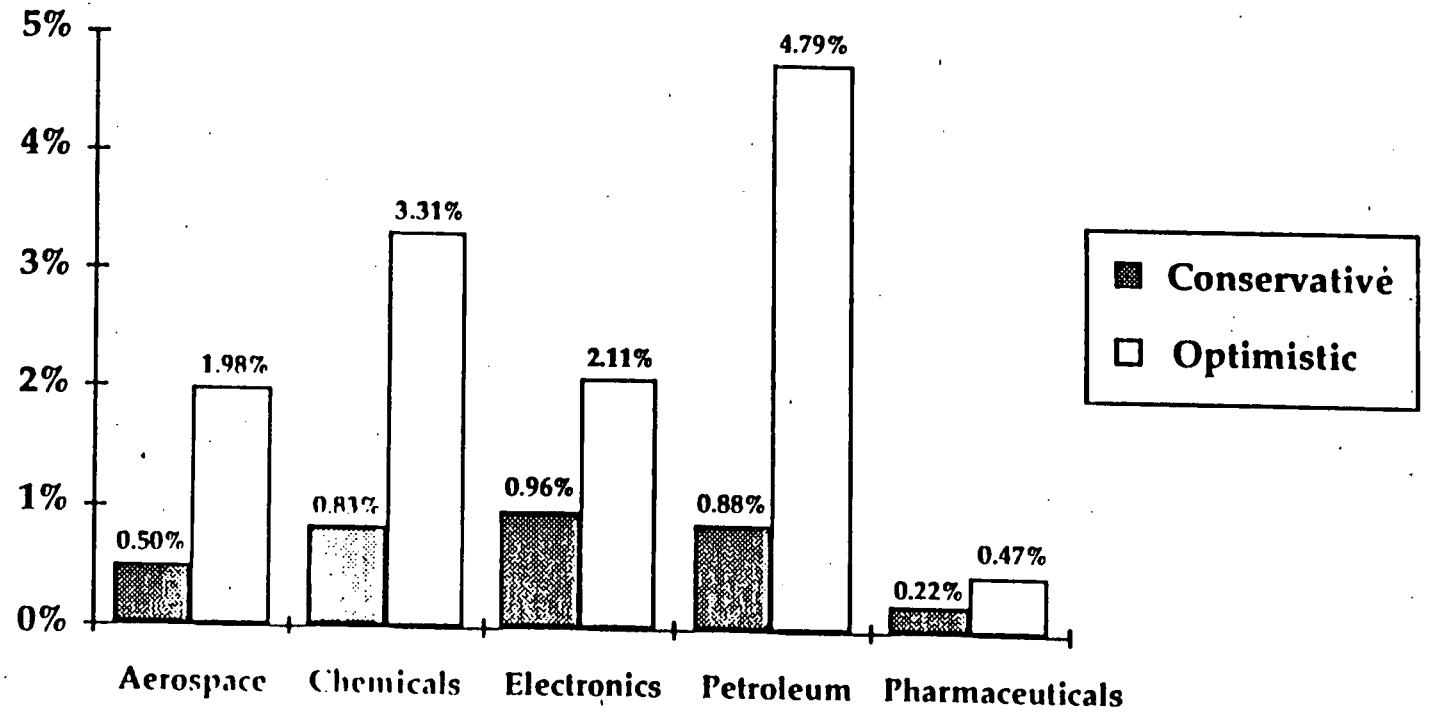
After translating the projected gains in R&D productivity to overall corporate productivity (based upon the ratio of R&D expenditures to total corporate spending), the variations in projected impact of the Federal HPCC Program in different industrial sectors appear to be related to the relative sophistication of these sectors in the application of HPC. Exhibit V-1 shows the expected increases in overall productivity over the coming decade in five industrial sectors.

Exhibit V-1: Projected Annual Productivity Increase, 1990-2000



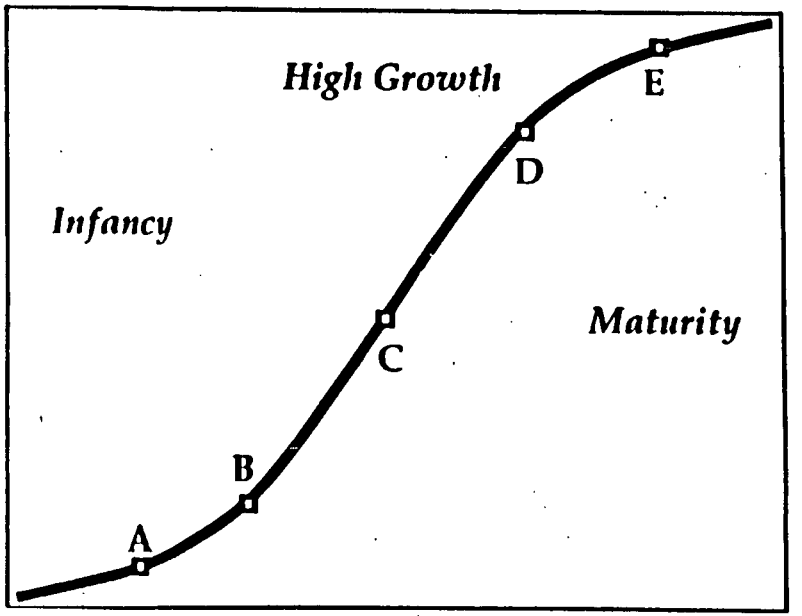
The expected impact of the Federal HPCC Program in each of these industrial sectors can be derived from Exhibit V-1 by subtracting the Scenario A estimates from the Scenario B estimates in each sector. The results are shown in Exhibit V-2.

Exhibit V-2: Annual Productivity Increases Resulting from HPCC Program, 1990-2000



The projected increase in R&D productivity is lowest in pharmaceuticals, which we take to be a consequence of the fact that pharmaceutical firms have only recently begun to invest heavily in HPC to support their research activity. Hence, we would put this industry at point "A" in Exhibit E-2 from Appendix E (duplicated below as Exhibit V-3), meaning that the really dramatic improvements are well into the future: that is, in the next century. This was also the opinion expressed by the R&D managers we surveyed in the pharmaceutical industry. (In the context of this study, we also feel that it is significant that the first pharmaceutical company to purchase a supercomputer, Eli Lilly & Company, did so as a consequence of its participation in the programs of the National Center for Supercomputing Applications at the University of Illinois. NCSA was established with Federal support, and its activities are precisely the kind which the proposed HPCC Program is intended to multiply and leverage.)

Exhibit V-3: Timing of Technological Investment

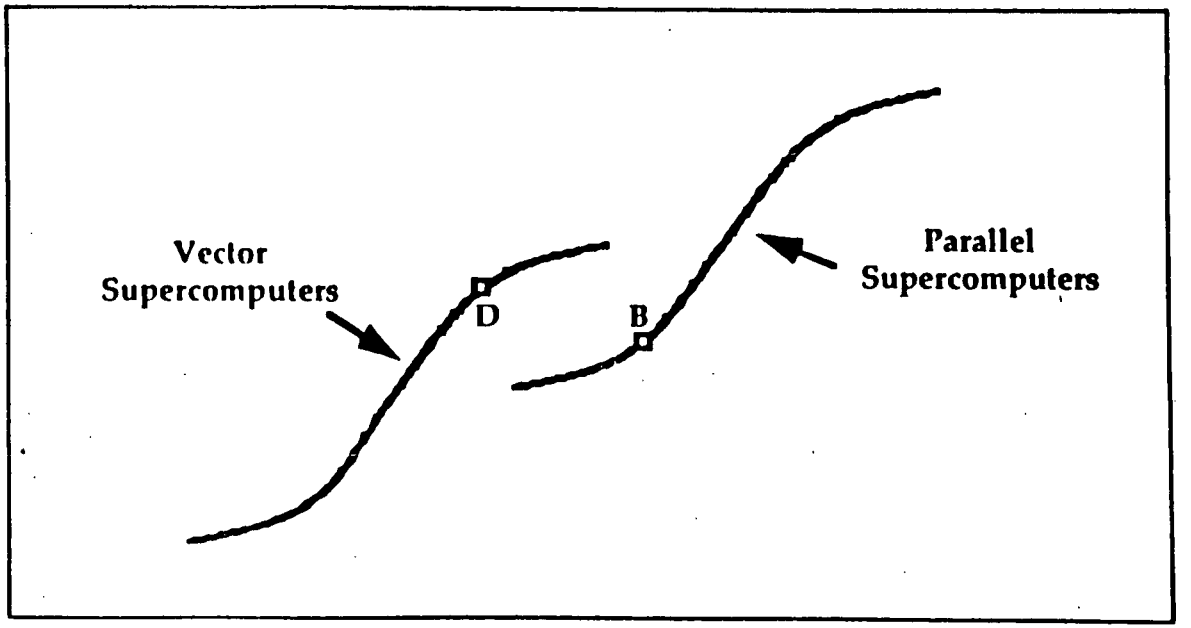


The chemical industry is next up the scale of sophistication in HPC application, perhaps at point "B," and this is reflected in the fairly ambitious growth expectations. After chemicals comes the electronics industry, which is somewhere around point "C." Thus, the optimistic projection of productivity increase in electronics is somewhat less than that for chemicals, but the conservative projection is greater. This pattern is repeated as we move on to the next sector, aerospace, which is also somewhere in the "C" to "D" range.

However, the interview results contain a "surprise" in petroleum. This sector, which rivals aerospace as the most sophisticated in its application of HPC, is projecting a rather large increase (relative to the other sectors) in R&D and overall productivity resulting from the Federal HPCC Program. The reason was evident in our discussions with the R&D managers: the petroleum companies are eagerly looking forward to exploiting the potential of highly-parallel HPC systems, probably more so than any other industry at present. They feel that they are approaching the point of diminishing return (point "D") on the technology curve for vector supercomputers, so they are preparing to jump to the "high growth" region (point "B") on the parallel computing technology curve (see Exhibit V-4). They expect that this will eventually enable them to develop applications which would be otherwise unattainable.

This illustrates another important benefit of the proposed Federal HPCC Program: helping users and vendors to identify promising new technologies (some of which, we trust, have also been nurtured through the "pre-competitive" stage with Federal support) and to migrate to them in a timely manner -- that is, as they are entering the "high growth" stage between points "B" and "D."

Exhibit V-4; Cascading Technology Curves

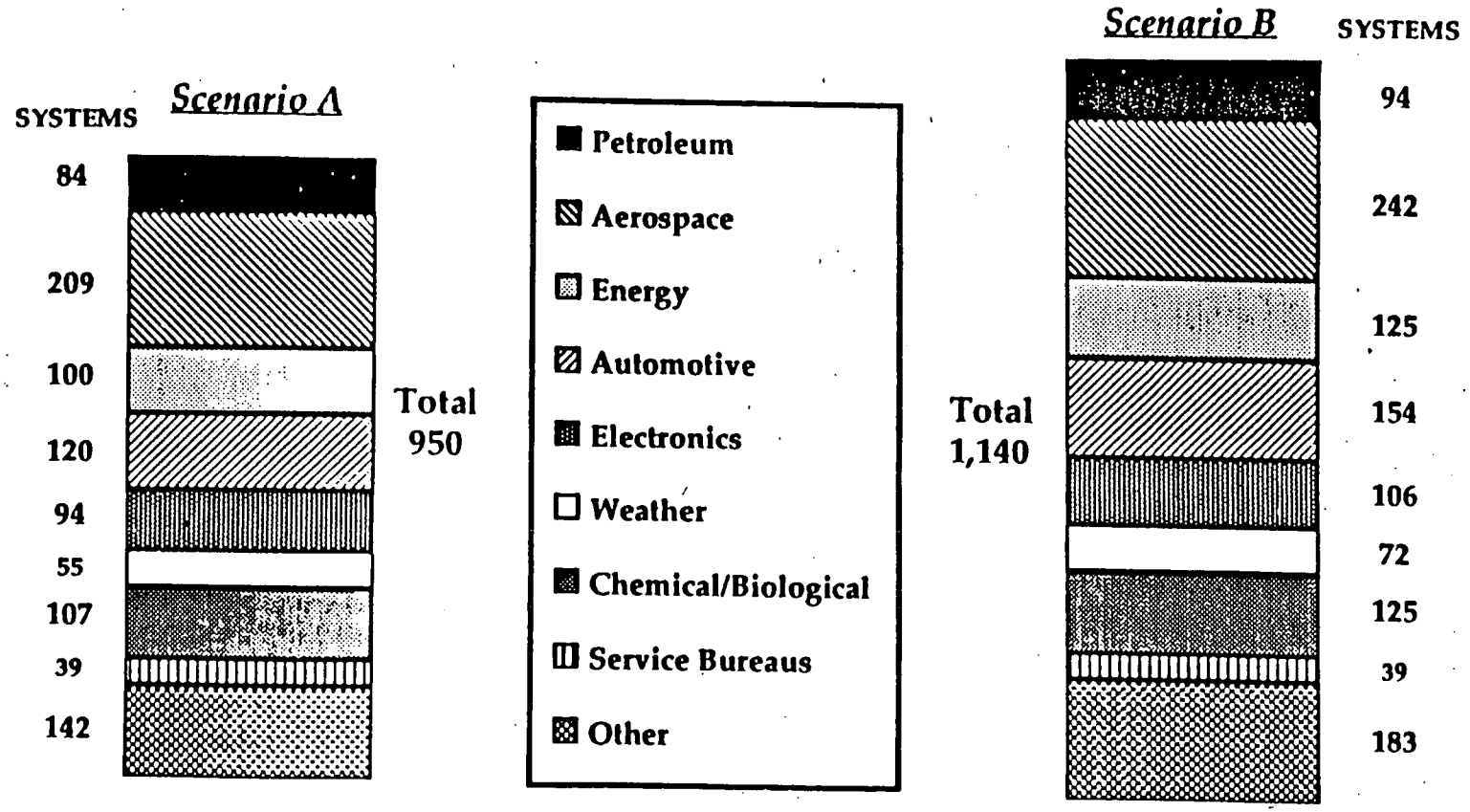


Although these projected productivity increases amount to only a few percent per year, their cumulative effect can be quite significant at the national level. Within the companies themselves, R&D is typically a highly-leveraged expenditure, which means that a relatively small improvement can have huge consequences, especially in today's intensely competitive markets. Also, because of the interlocking structure of various industrial sectors, a change in productivity in one sector can be quickly and deeply felt in several others.

Econometric models of the "input-output" type are especially suited to tracking these phenomena, which is why such a model was selected to extrapolate the above interview results to the U.S. economy as a whole. As described in Appendix B, the model used was the University of Maryland's Long-term Interindustry Forecasting Tool (LIFT), one of the most respected econometric models available. To drive the model, we estimated the impact of the Federal HPCC Program in all of the more than 60 economic sectors encompassed by that model, based upon our knowledge of the relative sophistication of HPC usage in these sectors and the results of our survey of the five industrial sectors identified above.

Exhibit V-5 shows the projected usage of supercomputers in various industrial sectors in the year 2000 under Scenarios A and B. As stated before, the principal impact of the HPCC Program will be upon the rate at which HPC permeates various applications, so the usage patterns under the two scenarios are quite similar. Where HPC usage is well-established and (therefore) comparatively sophisticated -- as, for example, in petroleum, aerospace, energy, and weather -- usage growth rates are relatively stable, so the impact of the HPCC Program is expected to be less dramatic than in applications areas where HPC usage has taken hold more recently: for example, automotive, electronics, and (especially) chemical and biological applications. Likewise, growth rates in supercomputer usage will be stronger in the latter applications areas under both scenarios. Where the HPCC Program is likely to have the greatest impact -- albeit perhaps not until after the year 2000 -- is in applications areas where HPC usage is presently very small, or even non-existent. Some possible examples are the food processing industry and service industries such as finance, transportation, construction, and entertainment.

Exhibit V-5: Year 2000 Installed Industrial Supercomputer Systems, Scenarios A and B



The projected productivity increases in the various industrial sectors were used to drive the econometric model. As described in Chapter IV for The HPC Arena, the impact of the Federal HPCC Program was determined by subtracting the model results for Scenario A from those for Scenario B. Actually, two Scenario B model runs were made: one using the "conservative" estimates of productivity improvement, the other using the "optimistic" estimates. Thus, we are fairly confident that the actual impact of the HPCC Program will fall within the range defined by these two estimates. (See also Appendix B for more details of the methodology.)

The principal result of this modeling study is this:

The proposed Federal HPCC Program will increase

the Gross National Product (GNP) of the United States by

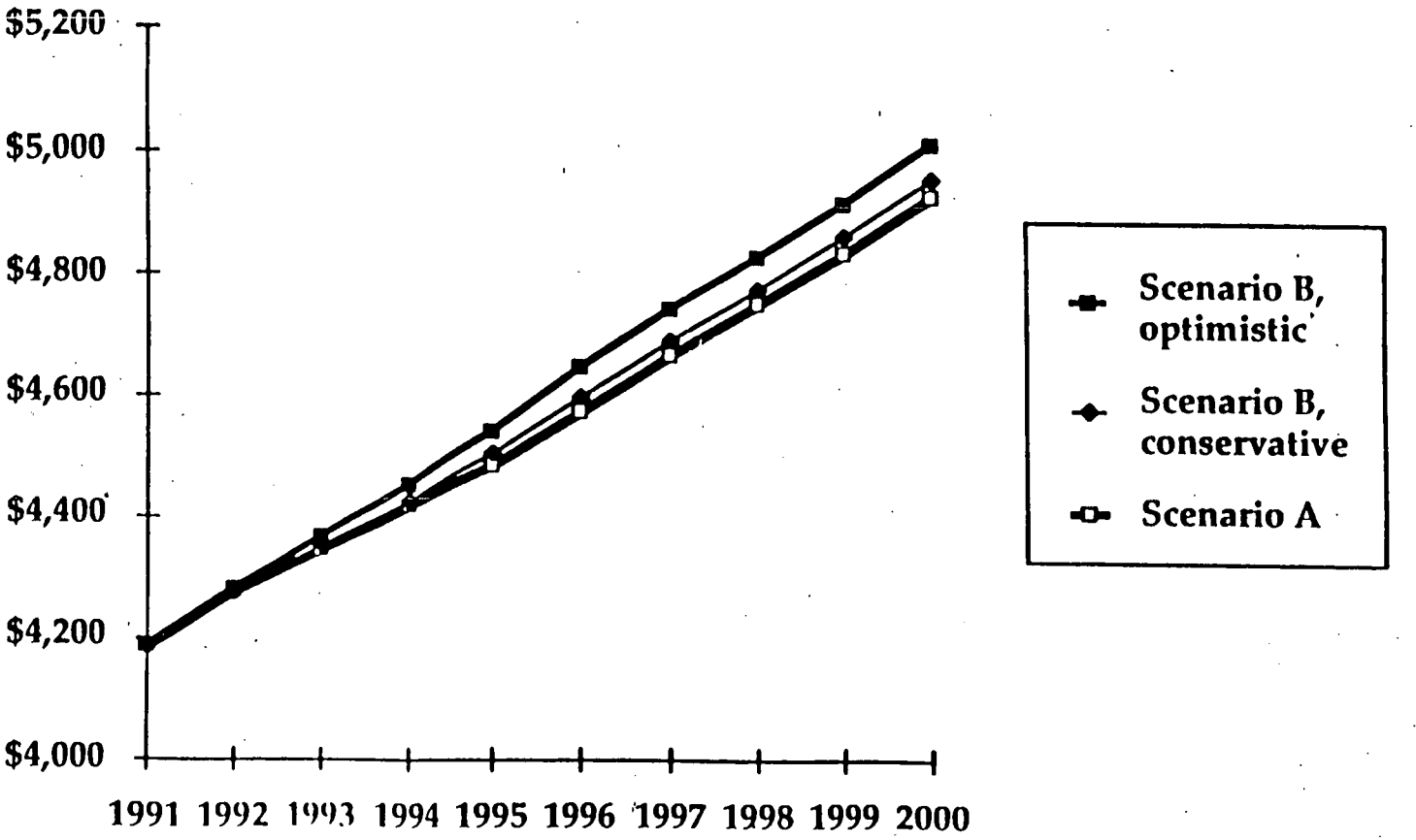
$172.5 to $502.6 billion

over the next decade.

This increase, which is given in 1982 "constant dollars," is relatively small in that it represents 0.4 to 1.1 percent of the GNP, but it is a very large payback -- at least, 138 to 1 -- on the investment. (In 1982 constant dollars, the $1.917 billion projected expenditure for the Federal HPCC Program is equivalent to about $1.25 billion.) Moreover, the difference between Scenarios A and B grows throughout the decade, which means that at the beginning of the 21st century, the U.S. GNP will be $28.9 to $83.9 billion (0.58% to 1.70%) greater if the Federal HPCC Program is funded (see Exhibit V-6).

Exhibit V-6: United States Gross National Product, 1991-2000

(1982 Constant Dollars, billions)



The model runs predict that other economic indicators will also be affected positively by the HPCC Program. For instance:

Over the next decade, the proposed Federal HPCC Program will:

- **Increase Personal Consumption by $101.8 to $280.6 billion;**
- **Increase Gross Private Domestic Investment by $57.5 to $199.2 billion;**
- **Increase Gross Exports by $8.4 to $30.6 billion and Net Exports (less Imports) by $3.2 to $22.8 billion;**
- **Lower average annual inflation by .016 to .044 percentage points; and**
- **Decrease the Federal deficit by $74.7 to $190.3 billion.**

(These numbers are all given in 1982 constant dollars.)

In the year 2000, the Federal HPCC Program would result in the following changes:

- **Personal Consumption: up $16.2 to $44.4 billion (0.54% to 1.47%);**
- **Gross Private Domestic Investment: up $8.5 to $25.7 billion(1.02% to 3.10%);**
- **Gross Exports: up $3.4 to $12.5 billion (0.34% - 1.25%);**
- **Imports: down $0.73 to $1.23 billion (0.10% - 0.16%);**
- **Net Exports: up $4.2 to $13.8 billion (1.68% - 5.65%); and**
- **Federal surplus: up $12.9 to $30.8 billion (27.38% to 65.26%) as a result of decreased government spending.**

(These numbers are all given in 1982 constant dollars.)

The only modeling result which might be interpreted as negative is that the Federal HPCC Program would increase unemployment by 0.11 to 0.40 percentage points (on a 4.67% base) in the year 2000. This represents 0.16 to 0.56 million people (6.49 million base). Obviously, this is a direct result of increased (by 0.78% to 2.40%) labor productivity in the private sector. However, the model predicts that the greatest impact on employment will be in the "Professional, technical, and related workers" category, where there is a chronic labor shortage (see Appendix H), so this is actually good news.

TECHNOLOGICAL LEADERSHIP

The projected impact of the Federal HPCC Program in terms of technological leadership is much harder to quantify than the productivity benefits. Nevertheless, we are confident that the effect on advanced technology will be quite substantial, and that this will have a significant salutary effect on the U.S. economy as well.

To provide a qualitative estimate of the technological benefits of the HPCC Program, we again use the paradigm of Exhibit V-3. For each of the "Application Opportunities" identified in Chapter III, Exhibit V-7 gives our assessment of the relative level of sophistication of HPC usage. This is expressed in terms of where, relative to points A, B, C, D, and E, the state-of-the-art is at present and where it will be in the year 2000. The benefit of the HPCC Program is the difference in the year 2000 sophistication levels for Scenarios A and B.

Exhibit V-7: Level of HPC Sophistication in Applications

	1990	2000	
		Scenario A	Scenario B
Materials Science	B-C	B-C	C
Semiconductor Design	B-C	C	D
Vehicle Dynamics	A-B	B	C
Transportation	C	C-D	D
Turbulence	B-C	B-C	C
Superconductivity	A	A-B	B
Efficiency of Combustion	B	B-C	C-D
Oil and Gas Recovery	C-D	D	E
Nuclear Fusion	B	B-C	C
Design of Pharmaceuticals	A-B	B-C	C-D
Structural Biology	B	B-C	C
Human Genome	A-B	B	B-C

Exhibit V-7 (cont'd)

	1990	2000	
		Scenario A	Scenario B
Prediction of Weather and Global Climate Change	C	C-D	D
Computational Ocean Sciences	C	C	C-D
Astronomy	B-C	B-C	C
Quantum Chromodynamics	B	B	B-C
Speech	B-C	B-C	C-D
Vision	B-C	C	D
Vehicle Signature	B-C	B-C	C
Undersea Surveillance	C	C	C-D
Engineering	B-C	C	D
Computational Chemistry	B-C	C	C-D
Film Animation	B-C	C	D
Bond Bidding	A	B	C-D

[This page has been left blank intentionally.]

CHAPTER VI - CONCLUSIONS AND RECOMMENDATIONS

VI - CONCLUSIONS AND RECOMMENDATIONS

This chapter consists of two sections, as follows:

- **Conclusions** - summarizes the principal findings of this study.
- **Recommendations** - presents some additional actions, over and above those contained in the proposed Federal HPCC program, which in our opinion would serve to strengthen and support that program.

CONCLUSIONS

"Prediction is very difficult, especially about the future," as (Nobel Laureate) Niels Bohr said. Nevertheless, we are confident that our experience in the information industry, coupled with that of the experts who have assisted us in this study (see Appendix C), has provided the best possible basis for projecting the course of High Performance Computing over the next decade.

If we follow "business as usual" (Scenario A), the future of HPC in the U.S. is uncertain at best. This is not to say that Japanese dominance in HPC is imminent, but the Japanese government, acting in concert with leading Japanese computer companies, has amply demonstrated its intention to insure a strong computer Japanese industry in general and a strong Japanese presence in HPC in particular. The Japanese government has also demonstrated its desire to encourage use of HPC in key industrial sectors, and if our analysis of the impact of HPC usage in the U.S. is anywhere close to correct, it is very likely that this is contributing substantially to the success of Japanese companies in worldwide markets.

On the other hand, if the United States invests in the proposed Federal HPCC Program (Scenario B), the future of the supercomputer industry looks considerably brighter -- 28 percent more revenue over the next decade -- although its well-being will still not be completely assured. Of greater overall significance to the country would be the marked expansion of HPC usage, which would enhance the productivity (and, hence, the competitiveness) of American industry. The projected increase in GNP is small in terms of percentages, but very large in terms of dollars.

In barest terms, we see the choice now before the nation to be a very simple one:

- **Either the U.S. risks loss of leadership in HPC, which we believe to be a highly significant and highly leveraged technological area (Scenario A);**
- **Or the U.S. takes the initiative to maintain and extend its leadership in HPC (Scenario B).**

Given the magnitude of the risk in Scenario A and the projected return-on-investment in Scenario B -- at least 138:1 and possibly as much as 400:1 -- we think that the proper path is clear:

The Federal HPCC Program, as proposed by OSTP, should be implemented as soon as possible.

In the words of Alan Kay (of Apple Computer): "The best way to predict the future is to invent it."

RECOMMENDATIONS

In the course of conducting this study, we have heard ideas and suggestions for strengthening U.S. activities in HPC, over and above those proposed in the Federal HPCC Program. Although our purpose in doing this study was not to elicit or evaluate such suggestions, we feel obliged to pass these ideas along to our Federal sponsors at this time, in the hope that they will be considered for incorporation in this important Program in order to further assure its success.

Technology Transfer

In our analysis for Scenario B, we assumed that the obstacles to HPC usage (described in Chapter III) will be overcome and that the "technology transfer" components of the HPCC Program will succeed in stimulating industrial demand for supercomputer systems. We regard these two activities, taken together, as the single most important element in attaining the overall goals of the HPCC Program. Indeed, one of the principal conclusions of the conference on "Frontiers of Supercomputing II: A National Reassessment" which was held at Los Alamos National Laboratory from August 20 to 24, 1990, was that:

"High Performance Computing, as a business,

will live or die according to its acceptance by private industry."

In this regard, we believe the Federal HPCC Program, as proposed by OSTP, to be necessary but not sufficient. In particular, we note that HPC usage is presently restricted to the very largest companies, *i.e.*, the "Forbes 500." Although the combined sales of these companies was $3.2 trillion in 1989, that represents only about 7.5 percent of corporate income in America. Clearly, if the U.S. is to realize the potential productivity gains from wider use of HPC, smaller firms must be made aware of HPC's benefits and assisted in learning how to utilize HPC's unique capabilities. For instance, it is not unreasonable to expect that a firm with $25 million or more in annual sales could benefit from using HPC. There are about 141,000 such firms in the U.S. They comprise less than 4 percent of the non-farm corporations, but their combined sales represents nearly two-thirds of the U.S. corporate total. They are also the sub-contractors and support infrastructure for the "top tier" of companies. Those are the companies that should be targeted by a Federal HPCC Program.

But how? Working with thousands of companies -- much less over 100,000 -- is clearly beyond the scope of the HPCC Program as it is presently framed. (This is not to criticize those who formulated the program. They did an admirable job within the boundaries, largely centering about R&D in HPC, they set for themselves. The point is that the Program should be framed more broadly.) Fortunately, there is a model for such an undertaking. It has been used in the United States for many years, and it has resulted in continuing productivity increases in the sector where it has been used. Consequently, the U.S. leads the world in that sector. The model is the Land Grant College system, established by the Morrill Act of 1862, plus the Farm Agent system, created by industrialist Julius Rosenwald in the early 1900s and later taken over by the government.

What we are suggesting is the creation of the industrial equivalent of the Land Grant/Farm Agent system to infuse HPC technology and techniques throughout American industry. The U.S. Department of Commerce (DoC) would be the logical agency to oversee this activity, and indeed, DoC has already instituted some activities of a similar nature in recent years as a result of legislation intended to bolster American competitiveness.

This would probably necessitate additional funding, well above the $1.917 billion now projected for the Federal HPCC Program. Parts of the existing HPCC Program might also have to be expanded if the proposed numbers of HPC centers and networks prove to be inadequate to serve this expanded user base. But the additional funds required would still be far below the potential payoff, and, without these additions to the HPCC initiative, the chances of attaining its ultimate goal -- namely, maximum American industrial competitiveness -- will be significantly diminished.

Alternative Education Programs

In our interviews of R&D managers, the shortage of trained personnel was frequently mentioned by companies who make and use supercomputers as the greatest barrier to the effective deployment and application of HPC in the United States (see Appendix H). And while the allocation of $183 million for "Basic Research and Human Resources" in the OSTP proposal was unanimously viewed as a positive step by everyone we interviewed, the general feeling was that it will be too little and too late. The country can scarcely wait for appreciation of computational science to seep down through the hierarchy of universities, colleges, junior colleges, and even high schools into the minds (and hearts) of future users and decision-makers. That will take at least two generations, by which time the competitive battles upon which our economic future depends will have become history.

However, in light of shrinking numbers of college-age persons in general and science/engineering majors in particular, it is doubtful that throwing more money at the problem will really help much. More drastic steps are needed to accelerate the transfer of knowledge from the relatively small number of experts in HPC to not only college students but also practicing scientists, engineers, and businesspersons. Hence, we suggest that it will be necessary to depart from "business as usual" in the educational system if the potential benefits from HPC are to be realized.

Of particular concern to us is the continuing and intensifying shortage of teachers in science and engineering, which has been exacerbated by the tendency of recent generations of students to eschew graduate study in favor of a quicker return on (educational) investment by taking industrial jobs immediately after earning the baccalaureate. (The late Robert Noyce, one of the founders of Intel Corporation, characterized this situation as "eating our seed corn.") Also of concern is the present pattern of limited interchange between academia and industry, in particular the fact that the flow of information about industrial practices and innovations into academia is largely indirect, except for the vacation employment of a few teachers in industrial laboratories and a relatively small band of industrial researchers who teach evening courses at a local college or university. We think that it is imperative, especially in HPC, that researchers from industrial and government laboratories be encouraged to return to teaching in order to maintain American scientific and engineering education at the highest possible quality.

Unfortunately, many would-be teachers in the private sector lack the necessary formal credentials -- for universities, an earned doctorate, and for public schools, teacher certification -- even though they certainly possess the expertise and the desire. Some states have moved to establish rapid certification programs for professionals switching into public school teaching, but for the non-PhD scientist or engineer wishing to teach at the college level the only options are accepting a sub-standard position (where the person's valuable expertise would be largely wasted) or returning to graduate school. In both of these cases, the financial and personal sacrifices are usually much too great, so the country's educational and technological infrastructure continues to erode.

The Japanese, however, do things differently. As in the U.S., nearly all Japanese students seeking an industrial career leave the universities after earning a bachelor's or master's degree. Only those who are planning to remain in academia as teachers or researchers pursue a doctoral course of study. But this does not mean that the others have lost forever the opportunity to earn a doctorate; many of them go on to earn a *rombun hakase* ("thesis doctor") degree from a Japanese university. (Japanese Ministry of Education statistics indicate that more than 60 percent of the doctorates awarded in Japan are of this type.)

Here's the way the *rombun hakase* system works. After establishing a reputation as a researcher in an industrial setting, a Japanese engineer or scientist can submit a thesis to the university of his (or, occasionally, her) choice. This need not be the *alma mater* where undergraduate study was done, but the thesis is not simply "thrown over the transom" either. Careful cultivation of relationships precedes the thesis submission, so that several "key" -- that is, senior and politically powerful -- members of the faculty are familiar with the candidate's qualifications. The thesis itself may consist simply of a collection -- three is usually enough -- of papers the candidate has published in internationally-refereed scholarly journals, or it may consist of a special research report prepared under partial supervision of some member of the faculty. In any event, if the thesis (and the candidate) are deemed worthy, the degree is granted forthwith -- no courses, no residency, no examinations, and only a modest fee paid as a courtesy to the professors who read and approve the thesis.

A significant number of Japanese industrial scientists and engineers who earn their doctorates in this way eventually return to academia -- typically, in their 40s or 50s -- as professors, and apparently their "status" is the same as those who completed traditional "course doctor" programs. (There is no easy way to tell which way a professor earned his degrees, and it is impolite to ask.)

This system has its origins in Europe, where similar practices may still exist, but the closest thing to it in the United States is found in the non-residential (and generally non-accredited) colleges which award credits toward degrees based upon students' "life experience." Despite long-standing efforts to give legitimacy and even accreditation to the more reputable of these institutions, most of them are regarded as little more than degree mills. But the fact that they exist at all gives testimony to the need for an alternative to formal classroom graduate education in America.

Hence, we suggest that established and accredited American universities adopt some version of the Japanese *rombun hakase* system. If this were to happen, a number of benefits would result:

(1) The number of persons with science/engineering doctorates could be quickly increased with essentially no change in the resources required. The quality and quantity of such doctorates, however, would remain under control of the universities.

(2) Academic-industrial technical exchange would become more of a two-way street, with quick and direct feedback in both directions.

(3) Industry would have an added incentive for building good relationships with local academic institutions, and the latter would gain broadened industrial support through increased numbers of loyal alumni.

(4) "Graduation fees" could even provide a modest financial windfall for universities.

The principal disincentive for establishing such a practice in America is resistance to change. Universities would have to openly admit that there are other, perhaps even better, ways of acquiring knowledge than sitting for a certain amount of time in a classroom. And given the notorious nature of university politics, such a radical step would be impossible unless there is a very strong external push behind it. To be specific, a modest amount Federal "seed money" should be allocated to convince a few "key" universities -- a handful whose reputation is beyond reproach -- to adopt the idea, after which others would probably cave in and follow their lead.

Another disincentive might lie in reluctance to imitate the Japanese, but that is a form of false pride that the U.S. can no longer afford.

Monitoring and Evaluation

To some extent, evaluation and feedback may be implicit in the envisioned management structure of the Federal HPCC Program, under which an HPC advisory panel would be formed to help OSTP and the FCCSET Committee on Computer Research and Applications monitor the progress of the Program. However, we believe that a full-time monitoring and evaluation effort is more appropriate, given the importance of this program to the nation's future. We also believe that these activities would be best carried out by an organization whose ties to the private sector are much stronger than its ties to the Federal bureaucracy*. This is because, as noted above, it is the private sector that the future of HPC depends upon.

In any event, the monitoring organization would conduct the field work and staffing which would support OSTP, FCCSET, and the HPC advisory panel in their oversight of the HPCC Program. Hence, it would strengthen, not replace, that oversight. An important element in this effort would also be continued econometric modeling in the manner used in the present study, to guide decisions regarding "mid-course adjustments" of the HPCC Program and to validate and further refine the modeling and forecasting techniques available to the Federal government for policy analysis.

* Of course, Gartner Group is such an organization, but that is not why we are making this suggestion. The relationships established and experience gained in the course of doing this study should be valuable to OSTP and FCCSET as the HPCC Program goes forward.

With many calculations, one can win; with few one cannot. How much less chance of victory has one who makes none at all!

-- Sun Tzu, **The Art of War**
(*circa* 500 B.C.)

[This page has been left blank intentionally.]

APPENDIX A - THE FEDERAL HPCC PROGRAM

[This page has been left blank intentionally.]

Exhibit A-1: HPCC Program Goals, Action Plans, and Funding

Component	High Performance Computing Systems	Advanced Software Technology & Algorithms	The National Research and Education Network (NREN)	Basic Research and Human Resources
Goals	Support the development of HPC systems capable of trillions of operations per second on significant problems.	Develop a base of software technology and algorithms that will: • Enable solution of Grand Challenge problems; • Have broad national impact on software productivity, capability, reliability.	Create a new NREN, operating at gigabits per second nationwide, within the next ten years: • Stage 1 -- Upgrade Internet to 1.5 Mbps; • Stage 2 -- 45 Mbps to 200/300 sites; • Stage 3 -- 1-3 Gbps to selected sites, expand 45 Mbps to 1000 sites.	Basic Research • Ensure adequate level of basic research to produce the next generation of innovative results in computing technology. Human Resources • Support basic research, education, and training. Support for Collaboration • Promote collaboration among the research community, industry, and government. Infrastructure • Provide facilities and research infrastructure, including hardware, networks, software, and application software.

... continued on next page

Exhibit A-1 (cont'd)

Component	High Performance Computing Systems	Advanced Software Technology & Algorithms	The National Research and Education Network (NREN)	Basic Research and Human Resources
Action Plans	Research for Future Generations of Computing • Increase research in computer science, scalable parallel computing, high density packaging, VLSI, and opto-electronics; • Develop components, packaging, and tools for large scale architectures. System Design Tools • Develop new generation of design tools and techniques: - Full cycle - Rapid prototyping. Transfer of Technology • Accelerate transition from (Federal) lab to market; • Pursue jointly high-risk projects. Evaluation of Early Systems • Basis for funding Software and Applications; • Grand Challenges fully weighed.	Support for Grand Challenges • Provide advanced software technology support to Grand Challenge researchers; • Shared facilities and testbeds on NREN. Software Components and Tools • Form collaborative groups to share software technology; • Provide incentives for industry to participate; • Develop (*e.g.*) distributed operating system for NREN. Computational Techniques • Support research in parallel computing algorithms; • Develop higher level languages for computational scientists. HPC Research Centers • Support deployment of HPC architectures to Grand Challenge researchers; • Provide facilities to computing technology researchers.	Interagency Effort • Coordinate diverse agencies to foster support; • Enhance network security. R&D for Gbps Net • Define structure of Stage 3 network; • Develop new switching systems and protocols. Deployment of Gbps NREN • Mid to late 1990s. Structured Transition to Commercial Service • Process for transition of NREN from government operation to a commercial service.	Basic Research • Expand basic research; • Provide NREN access to all; • Improve facilities available for research and education. Human Resources • 1,000 Ph.D.s per year by 1995; • Promote 10 degree programs; • Upgrade 10 university computer science departments; • Improve ties between computer technology and other disciplines; • Provide access to professional engineering support.

... continued on next page

Exhibit A-1 (cont'd)

Component	High Performance Computing Systems	Advanced Software Technology & Algorithms	The National Research and Education Network (NREN)	Basic Research and Human Resources
Funding (five years)	$682 Million	$662 Million	$390 Million	$183 Million
		Total Funding: $1.917 Billion		

[This page has been left blank intentionally.]

APPENDIX B - RESEARCH METHODOLOGY

[This page has been left blank intentionally.]

OVERALL APPROACH

The purpose of this Gartner Group study is to develop a quantitative assessment of the likely economic impact of the proposed Federal HPCC Program over the coming decade.

This study has been conducted in two phases:

- In Phase I, two alternative scenarios, depicting supercomputing through the year 2000, were developed. One scenario assumes full funding for the proposed HPCC Program, commencing in FY 1992; the other scenario assumes "business as usual" -- that is, no additional Federal funding above what is expected for HPC-related activities now underway.
- In Phase II, these scenarios were extended to encompass the impact of HPC, first upon selected industrial segments which are major users of supercomputers and then upon the U.S. economy as a whole.

PHASE I APPROACH

Gartner Group carried out the Phase I study in approximately eight steps, as follows:

1. **Brainstorming** - Initially, we convened a brainstorming session with selected Gartner Group analysts having particular expertise in High Performance Computing (HPC). After reviewing the government's proposed HPCC Program and the objectives of this study, we formulated a framework for the alternative scenarios and an agenda for research.

2. **Literature Search And Review** - We then searched for and reviewed various pertinent documents, including government-sponsored papers, trade and general news articles, and Gartner Group research notes.

3. **Client Review** - We met with representatives of DOE, LANL, and OSTP to review a preliminary outline of the report and to discuss issues related to the study.

4. **Interviews Of Subject Experts** - Next, we conducted interviews with various members of the vendor, academic, and research communities. These interviews focused on technical and applications bottlenecks and the impact of removing these bottlenecks.

5. **Construction of Background Scenario** - Using the information obtained in the previous steps and our knowledge of the history and dynamics of the information industry, we prepared a background scenario which traces HPC development from 1980 through the present.

6. **Formulation of Scenarios** - Based upon the insights gained from the background scenario and again drawing upon information gleaned in the previous steps, our understanding of the international information industry, and other Gartner Group scenarios already on hand, we prepared the alternative scenarios for the years 1990-2000.

7. **Peer Review** - The scenarios were then reviewed by experienced Gartner Group analysts for reasonableness and consistency with the overall computer industry scenarios which are regularly updated by the company in the course of its ongoing research process.

8. **Report Preparation** - A preliminary report was then prepared and circulated to a limited number of experts in the government and the supercomputing industry.

Over the last three years, Gartner Group has developed a quantitative model which characterizes the information industry in terms of MIPS, systems, and dollars for various classes of systems -- mainframes, minicomputers, personal computers, etc. -- as well as the components of these systems: CPUs, peripherals, software, etc. Both the methodology and the results of this model have been applied in Step 6 to develop the two alternative scenarios for the coming decade in supercomputing.

Basically, the Gartner Group information industry model assumes that technology is the driver of demand because it is the principal determiner of both the overall performance and the price/performance of various types of information systems. Hence, future projections are based upon anticipated technological advances, interpreted through our understanding of the effects of similar advances in the past and of the changing competitive conditions in the industry. Industry revenues are derived from these projections of price/performance and of MIPS and systems shipments, using average system price and average MIPS per system as "reasonability" checks. (Historically, the model also reflects macroeconomic cycles which have affected overall demand, but there has been no attempt to incorporate macroeconomic forecasts into the future projections except for one factor: all revenue, price, and price/performance figures are given in "current" dollars. Because these "current" dollar amounts reflect inflation -- averaging about 3% per year -- over the past decade, forward dollar projections are also similarly biased.)

Thus, our projection of the future of supercomputing is rooted in our understanding of the past, not only of supercomputing but also of other elements of the information industry.

PHASE II APPROACH

Our fundamental assumption is that supercomputers find their primary usage in research and development (R&D), as an adjunct to, and partial replacement for, laboratory or field experimentation and testing. Thus, HPC enables companies to

- Bring more and better new products to market; and
- Bring new products to market faster.

In other words, HPC improves R&D productivity. Even if there is no other benefit from the use of HPC, this change in R&D productivity affects overall company productivity in direct proportion to the share of expenditures for R&D. This then forms the basis for our approach:

1. **Expert Survey** - Scenarios A and B from Phase I were presented (in abbreviated form) to company R&D managers from five industrial sectors, representing a variety of experience and sophistication in the application of HPC:
   - Aerospace;
   - Chemicals
   - Electronics
   - Oil & gas exploration and production; and
   - Pharmaceuticals.

   For both Scenarios, these managers were then asked to give "conservative" and "optimistic" estimates, based upon their expertise and experience, of the change in R&D productivity in their respective companies over the coming decade.

2. **Data Analysis** - For both Scenarios, these estimates were translated into overall productivity improvement estimates for the companies, using information (from their annual reports) about the ratio of R&D spending to total spending. Next, productivity estimates for several companies in the same industrial sector were combined, with weightings based upon relative revenues, to obtain "conservative" and "optimistic" overall Scenario A and B productivity improvement estimates for the five industrial sectors identified above. Finally, corresponding productivity improvement estimates were made for other industrial sectors, based upon their relative sophistication and rate of change in the utilization of HPC in their R&D.

3. **Simulation** - The results of Step 2 were used to drive an input-output econometric model, the Long-term Interindustry Forecasting Tool (LIFT) of the INFORUM research group of the University of Maryland at College Park. This is an integrated macroeconomic model that provides detailed projections of industrial production, employment, and prices, as well as personal consumption expenditures, equipment investment, construction investment, government purchases, imports and exports over 15-20 year horizons. It the culmination of more than 20 years of research and development by Professor Clopper Almon and is now recognized as one of the best economic forecasting tools available. It has been used for various economic studies by government agencies such as the U.S. Departments of Commerce and Labor, the Federal Emergency Management Agency, the Congressional Budget Office, and the Japan External Trade Organization (JETRO). In our usage of this model, Scenario A data were taken as the "base" case, and Scenario B data as the "incremental" case. The difference between these two cases was taken to be the economic impact of the proposed Federal HPCC Program. Separate runs were made for "conservative" and "optimistic" estimates to obtain an impact range.

4. **Report Preparation** - This final report was then prepared. Comments received from review of the Phase I preliminary report were used in revising material from that report as it was incorporated into this one.

SCENARIOS

The **Random House Dictionary** defines "scenario" as "an outline of the plot of a dramatic work, giving particulars as to the scenes, characters and situations." Gartner Group, Inc. has made a business of applying the scenario concept, perhaps with a bit of poetic license, to the information industry. Thus, we are constantly developing and refining our vision of the evolving drama of the information industry, giving particulars as to the technologies, the vendors, the markets, and the users. We build our scenarios both "bottom up" and "top down," in an iterative process, drawing upon the wisdom, experience, and insight of our cadre of analysts, who collectively represent several hundred years of experience in the industry.

These same methods, sources, and analysts were used to develop the scenarios which follow. Of course, they are not infallible -- we do not claim to have precise views of the year 2000 -- but they were, in our judgment, the best available. They included:

- **Existing Gartner Group Scenarios:** Gartner Group provides continuous research on 14 sectors of the information industry to over 800 companies and government agencies. These scenarios are fundamental to our research process and were used to develop Scenarios A and B below.
- **The Gartner Group Information Industry Model:** Gartner Group has developed a quantitative model of the information industry as a basis for its scenarios. Consisting of a number of interlinked spreadsheets containing data on unit shipments and revenue for various industry segments over a 20-year period, the model can be used to identify trends and test assumptions about the dynamics of the information industry.

... continued on next page

- **Historical Background:** Developments from 1980 to 1990 were used to provide a partial, extrapolative basis for Scenarios A and B (see Chapter IV). In so doing, we gave more weight to recent events, such as Control Data Corporation's withdrawal from the supercomputer business, and to trends which we feel will dominate the coming decade, such as the ascendancy of Japanese supercomputer companies.

- **Assumptions:** It was assumed that Federal funds will be spent wisely: that is, that Federal managers will make the best possible use of the resources as events unfold and that intervention in the private sector will be minimal consistent with the overall goals of the program.

- **Experience:** The principal consultants who prepared this report each have more than 30 years experience, both domestic and international, in the information industry, including work on various aspects of High Performance Computing, spanning academia, government, and the private sector.

- **Peer Review:** After the scenarios were drafted, they were reviewed by other Gartner Group experts with experience in High Performance Computing. The scenarios were then published in the Phase I preliminary report, which was circulated to selected experts in the Federal government and in the the HPC community. Comments and criticisms received were used in revising the scenarios for the draft version of the final report. Again, copies were circulated to selected experts, and their comments were used in the preparation of this document.

To put a label on it, the technique used is best described as "jury of expert opinion."

[This page has been left blank intentionally.]

APPENDIX C - SOURCES

[This page has been left blank intentionally.]

SOURCES

The following persons contributed to this study as advisors, critics, reviewers, and sources of data and/or expert opinion, and their support is herewith gratefully acknowledged.

Active Memory Technology, Inc.	Geoff Manning
Amoco Corporation	J. F. (Joe) Gentile and Keith McHenry
BBN Advanced Computers	Ben Barker
The Boeing Company	Philip H. Nelson, Stephen Niver, and Melvin Scott
Brett Berlin Associates	F. Brett Berlin
Chevron Corporation	William Bartz and J. R. (Rex) Bell
Ciba-Geigy Corporation	Jerald R. Paules and Mike Scott
Control Data Corporation	Kent Steiner
Convex Computer Corporation	Bob Paluck and Steve Wallach
Cornell University	Alan McAdams

SOURCES (cont'd)

Corporation for National Research Initiatives	Robert E. Kahn
Cray Computer Corporation	Neil Davenport
Cray Research, Inc.	Bob Ewald and staff, Ed Masi and staff, and John Rollwagen
DataMax, Inc.	Lloyd M. Thorndyke
Defense Advanced Research Projects Agency	Stephen L. Squires
Digital Equipment Corporation	Samuel Fuller and Walter Kasell
E. I. du Pont de Nemours and Company	David A. Dixon and Howard Simons
FMC Corporation	John Luranc
General Motors Research Laboratories	Myron Ginsberg
HNSX Supercomputers	Samuel Adams
Intel Corporation	Avram Miller, Gerhard Parker, Justin Rattner, and Donald L. Scharfetter
Eli Lilly and Company	John S. Wold

SOURCES (cont'd)

Lawrence Livermore National Laboratory	Robert Borchers and George Michael
Los Alamos National Laboratory	Norman Morse and W. L. (Buck) Thompson
Martin Marietta Corporation	Norman R. Augustine, David A. Dieterich, Bert Westwood, and Raymond S. Wiltshire
McKinsey & Company	Lawrence H. Linden
Merck & Co., Inc.	Myra Williams
Mobil Corporation	Norman Guinzy and W. R. (Bill) Rhodes
Monsanto Company	William Fleming, Bipin Junnarkar, Robert E. Otto, and John C. Schaefer
Motorola Inc.	Bernard W. (Buck) Jordan, Jr.
Multiflow Computer Inc.	Donald E. Eckdahl
National Aeronautics and Space Administration	Lee B. Holcomb
National Research Council	Marjorie Blumenthal and Martha Harris
National Security Agency	Norman S. Glick and Kermith Speierman
National Science Foundation	Charles Brownstein, Melvyn Ciment, Susan T. Hill, and Stephen S. Wolff

SOURCES (cont'd)

Office of Naval Research	David K. Kahaner
Shell Development Company	Patrick Savage
Supercomputing Research Center	Alfred E. Brenner, Richard N. Draper, John P. Riganati, and Paul B. Schneck
Supercomputer Systems, Inc.	Leslie Chow
Thinking Machines Corporation	Jim Bailey, Richard Clayton, Sheryl Handler, Danny Hillis, and David Orr
United Technologies Corporation	Robert J. Hermann
University of Arizona	Seymour E. Goodman
University of California, Berkeley	Michael A. Harrison
University of Houston	John Killough
University of Illinois - Urbana/Champaign	Jim Bottum, Lloyd C. Hodges, Larry L. Smarr, and John Stevenson
University of Maryland, College Park	Clopper Almon, Margaret McCarthy, Lorraine Monaco, and Doug Nyhus
University of South Carolina	Paul Huray

SOURCES (cont'd)

University of Washington	Marie Anchordoguy
U.S. Department of Commerce	James Burrows, John McPhee, Jonathan Streeter, and Robert White
U.S. Department of Energy	John Cavallini, Norman H. Kreisman, and David B. Nelson
U.S. Embassy, Tokyo	Michael V. McCabe and Seikoh Sakiyama
U.S. Senate Staff	Michael R. Nelson
Worlton & Associates	Jack Worlton
Others	C. Gordon Bell
	Sidney Fernbach
	Hiroshi Mizuta
	Yoshio Shimamoto

Last, but certainly not least, Dr. Herbert E. Striner, a consultant to Los Alamos National Laboratory, served as principal advisor -- intellectually, emotionally, and spiritually -- for the entire study. To the extent that the study succeeds in meeting its objectives, he deserves a large part of the credit; to the extent that it does not, he is hereby absolved from any blame.

[This page has been left blank intentionally.]

APPENDIX D - HPC CONCEPTS

[This page has been left blank intentionally.]

HPC TERMINOLOGY

The elements which comprise HPC systems are:

Processors: Processors perform the basic computations in HPC systems.

Memory: Central memory is where data are stored during problem solution. In "von Neumann" computers, it is also where the control programs are stored. Hence, the size of the memory determines the maximum size of the problems which can be solved, and the speed of the memory limits the speed of computation.

Peripherals: Peripheral devices are used for input and output of information. Often, they also provide auxiliary storage of data and programs.

Networks: Networks allow the rapid exchange of information between HPC systems and/or between system components: *e.g.*, between workstations and supercomputers.

Workstations: Workstations provide the human interface to HPC. Usually, they are rather powerful computers dedicated to controlling output devices such as graphics displays, etc.

Programs: Programs provide the control information to tell HPC systems what to do.

Algorithms: Algorithms are the abstract problem-solving methods which are encoded in programs.

Applications: Applications are the uses to which HPC is put, the problems which are to be solved.

HPC COMPONENTS

Components are the elementary building blocks of HPC systems, so system improvement can be realized through technological advances in semiconductor circuits (including techniques for designing and building them), magnetic devices and materials, communications network components, display screens, etc.

Until about 1970, silicon (*Si*) was assumed to be the basic material for all computer circuits. As early as the 1950s, however, compounds with greater electron mobility -- in particular, gallium arsenide (*GaAs*) -- were being studied as the likely semiconductor materials of the future. The practical use of *GaAs* has been slow in coming, because its characteristics are different enough from silicon that it presents new and difficult manufacturing challenges. Hence, silicon is still the dominant material at present, with gallium arsenide about to make a debut. Convex will use *GaAs* in its C3 supercomputer series, expected in 1H91; Seymour Cray will also use *GaAs* in his Cray-3 system, due out in late 1991 or early 1992; and the Japanese vendors are expected to announce *GaAs*-based supercomputers, perhaps in 1992 or 1993, as a result of *GaAs* R&D programs they have had underway for some time.

Attempts to achieve greater speed through superconductivity have also been explored -- and largely abandoned because of the problems associated with the extremely low temperatures required -- but the recent discovery of high-temperature superconductivity may eventually provide longed-for breakthroughs. However, this technology will probably not be ready for commercial use in computers until the 21st century. It is also likely that other technologies which did not even exist ten years ago (and some of which may be almost unknown today) may emerge to play important roles in HPC by the year 2000. Some possible candidates, which are now being researched in the U.S. and abroad, include optical circuitry and high-density biocircuits made from organic compounds.

HPC ARCHITECTURES

HPC systems can also be improved through advances in their architectures: that is, the ways in which the components are utilized, their interrelationships, etc. This includes techniques for HPC system design, construction, test, and maintenance. As component technologies have approached certain fundamental physical limitations -- *e.g.*, the inability to reduce line size on semiconductor chips below a few atoms wide -- architectural innovations have become increasingly important in improving overall computing performance.

The basic characteristic of computers is that they execute a collection of instructions (called a program) in sequence. Hence, their speed has traditionally been measured in terms of how many millions of instructions per second (MIPS) they can perform. (This measurement is fraught with difficulties, because not all instructions do the same amount of work, but nevertheless MIPS is widely used as a crude approximation of computing power.)

Whereas component improvements tend to affect directly how many MIPS a computer of any given design can execute, architectural innovations effectively increase MIPS by more subtle (and complex) strategies. If we liken instruction execution to building an automobile, component improvements are the analogue of working harder -- that is, faster -- while architectural improvements equate with working smarter. For instance, instead of building just one car at a time, some workmen might be assigned the job of gathering the parts for a second car while their colleagues are completing the assembly of the first one. By analogy, at the same time a computer is executing one instruction, it can also be fetching the next one (from memory) -- unless, of course, the first instruction can affect what the "next" one will be. This is called **instruction overlap**. It is a technique first used in supercomputers 30 years ago, but it is now commonplace even in desktop personal computers. (High performance computer systems have frequently been the proving ground for components and architectural approaches which have subsequently been employed throughout the industry.)

A more complex version of the same idea is to begin the execution of the second instruction before the first is completed -- again, if there are no complicating interdependencies between them. This is called **pipelining**. It is analogous to an assembly line, and like an assembly line, the concept can be extended to permit several successive instructions to be in different stages of execution simultaneously. This obviously requires a more complicated and expensive computer design (or factory set-up), but the resultant impact on MIPS (or auto production) is generally regarded as worth the effort.

The next architectural innovation in HPC, however, involves re-stating the problem. In terms of the automobile analogy, the problem becomes one of "capacity to transport people," rather than "production of cars" -- and the solution is to build some buses. Thus, instead of operating only upon scalars (which might be likened to single-passenger cars), some supercomputers can also perform vector operations -- in effect, a whole "busload" at a time. (Mathematically, a vector is a row or a column of numbers.) For instance, if A is a row vector of five elements, expressed:

$$A = (a_1, a_2, a_3, a_4, a_5)$$

and B is another row vector of five elements:

$$B = (b_1, b_2, b_3, b_4, b_5)$$

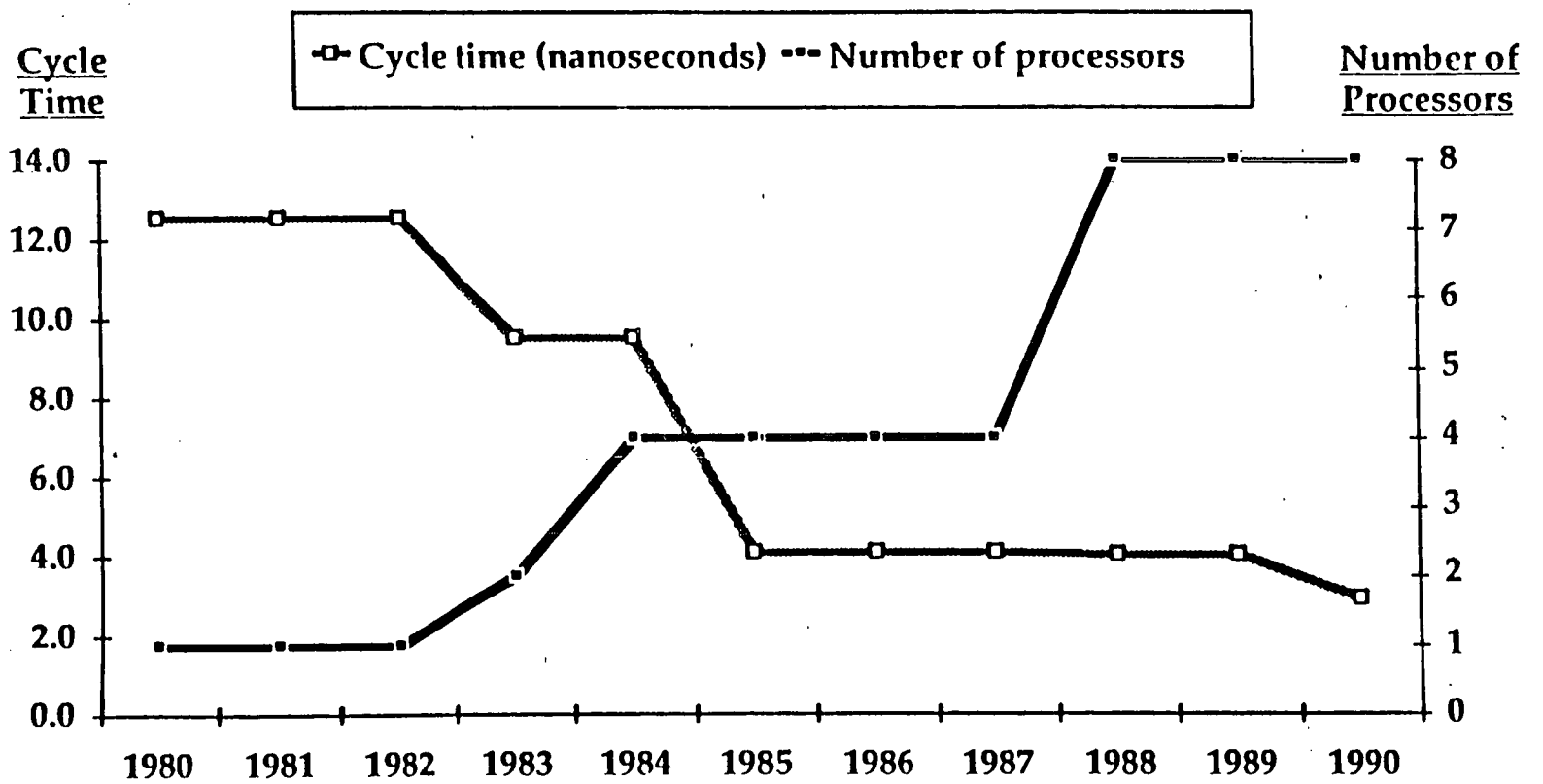
then the vectors A and B can be "added" by adding their respective elements:

$$A + B = (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4, a_5 + b_5)$$

This concept can be extended to vectors of any length (but vector addition "works" only if the vectors are of the same length), and of course, the addition can be "pipelined" (as it usually is). The advantage is that one vector instruction does the work of several scalar instructions (which is the principle behind car-pooling). Again, this is much more complicated and costly than "scalar" processing, but if maximum performance is of the essence, the cost is justified. However, because a vector instruction does so much more work than a scalar instruction, it would be stretching things to measure vector performance in MIPS. So, vector performance is measured in terms of how fast the element-by-element processing of vectors is done: that is, the rate at which results emerge from the (floating-point) pipeline, millions of floating-point operations per second or "megaflops" (MFLOPS). Hence, MIPS is used as a measure of scalar processing speed, and MFLOPS as a measure of vector processing speed.

The final architectural approach to achieving greater performance is **parallelism**. In terms of our analogy, this means simply building multiple factories (multiprocessing), or putting multiple assembly lines side-by-side within the same factory (multiple pipelines). Note that the processing can be either scalar or vector, but the performance is usually expressed in terms of megaflops. Parallelism, however, increases hardware cost and complexity and tends to reduce reliability, because of significantly greater component counts. Nevertheless, as advances in semiconductor switching speeds become more difficult to achieve, increased use of parallelism is viewed as inevitable, even in vector supercomputers (see Exhibit D-1).

Exhibit D-1: Vector Supercomputer Characteristics



Parallelism can take a number of forms, depending upon the level (or degree) and type. Any computer system with two or more processors is, strictly speaking, a parallel system, but it is not until the level exceeds 64 processors that we speak of **large-scale parallelism**. In between, the term **medium-scale parallelism** is applied, especially to systems with 16 to 64 processors -- the bounds tend to shift upward over time -- and **massive parallelism** characterizes very large systems, with 1,024 or more processors. In this report, the term "highly-parallel" is used to denote large-scale and/or massive parallelism.

The two principal types of parallelism are SIMD and MIMD. In SIMD -- Single-Instruction-Multiple-Datastream -- parallelism, all processors simultaneously perform the same instruction on different data elements. This approach (which is used in the Connection Machine, for example) simplifies the programming of a large number of processors, but it usually requires that algorithms and software for vector supercomputers be re-thought and re-written -- that is, "parallelized" -- in order to realize the potential performance increases built into the hardware. Moreover, some experts believe that there may be significant problems which are inherently "un-parallelizable" and which, therefore, are inappropriate for solution on SIMD systems.

MIMD -- Multiple-Instruction-Multiple-Datastream -- architectures (which are used in Intel's Hypercube, for example) offer greater flexibility than SIMD systems, but they also present the added problem of coordination and synchronization between two or more processors working on different parts of the same problem. Hence, their development, as reflected in the number of processors operating in parallel, is likely to lag behind that of SIMD systems, at least in the near term. However, gains are expected in their usage, as a result of software and mathematical research, and in their processing power, as a result of hardware and architecture development, over the next decade, because it is to this variant of parallelism that Cray Research and IBM seem to be evolving in extending their present product lines.

As should be obvious from the foregoing description, programming for vector computation is quite different than programming for scalar computation. Although "vectorizing" compilers have been developed to convert programs written for scalar computers for use on vector computers, inefficiencies usually result. The same holds true, but perhaps in spades, for parallelism: inter-processor communication (or the lack thereof) can seriously degrade performance. But parallelism is apparently inevitable in the future of computing, as it becomes increasingly difficult to build faster single processor systems. Hence, it will be necessary to devise new programming methods and new algorithms to fit the new architectures.

HPC SOFTWARE

Software (and the lack thereof) is often a major barrier to computer usage of any kind, and HPC is certainly no exception. Whereas the small cadres of early supercomputer users were able to make do with rudimentary software systems and tools and were frequently willing to develop sophisticated software packages to fill in the voids in vendor-supplied software, the average computer user of today is unable – or, at least, unwilling -- to do so. And although much progress has been made by supercomputer vendors and independent software vendors in the past decade, HPC systems are still among the least "user-friendly" of those currently available. This problem is exacerbated by the growing need for HPC users to "think parallel" and generally break out of the mental straightjacket of "traditional" scalar programming.

As a result of the complexity of HPC software and the relatively small size of the market, the prices of HPC software packages tend to be high. Market fragmentation has also compounded this problem: software for one type of computer system usually won't work on another type. To help remedy this situation, there has been a definite trend toward universal adoption of the UNIX operating system for supercomputers, minisupercomputers, and workstations in recent years. But UNIX itself is not especially user-friendly either, and it has some notorious shortcomings which must be overcome to make it more suitable for HPC environments. Moreover, while UNIX standardization may facilitate the portability of applications programs between dissimilar systems, the need to "fine tune" programs to suit architectural peculiarities is still a significant barrier to achieving maximum performance.

HPC PERFORMANCE

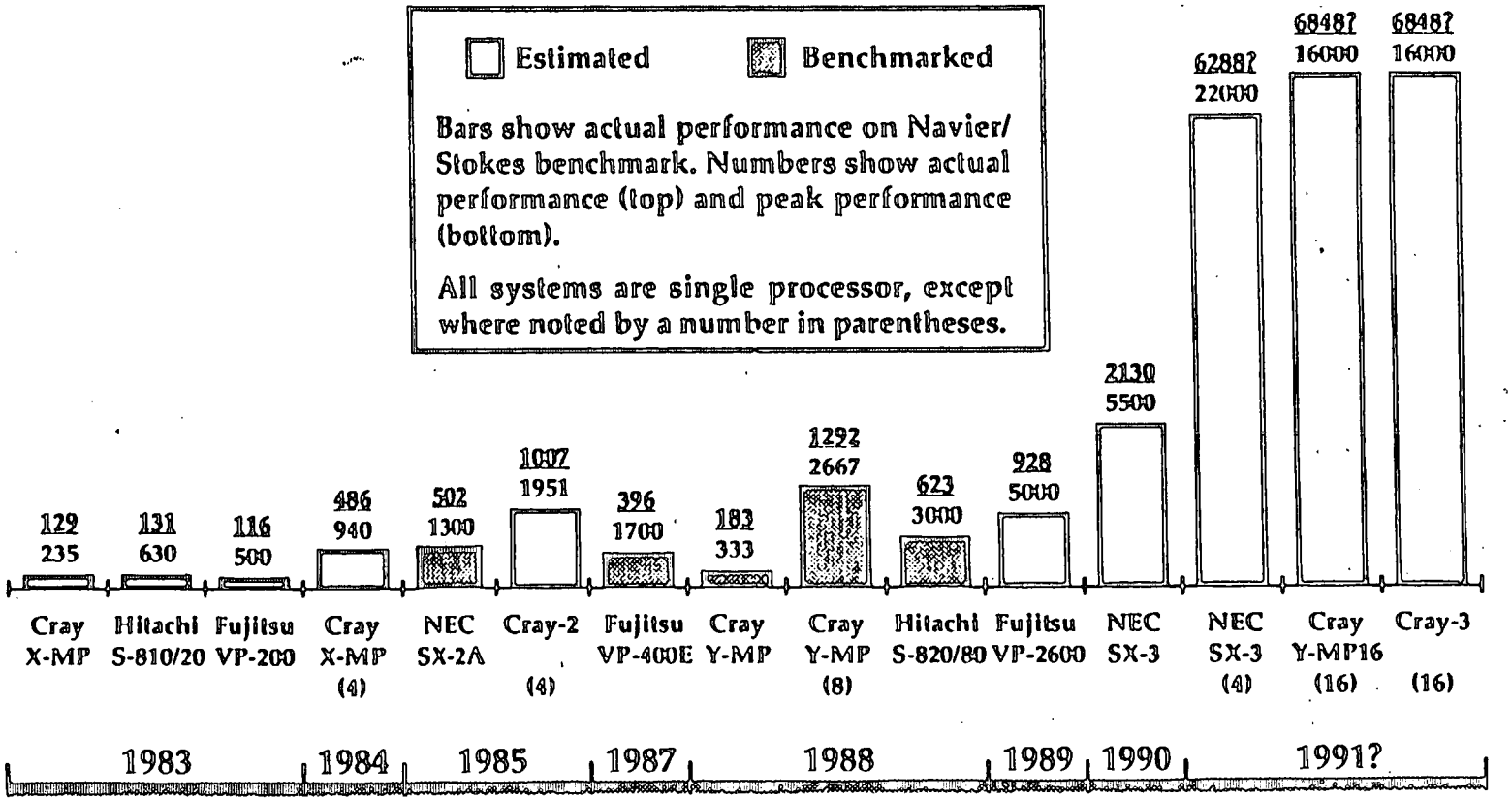
Perhaps no aspect of supercomputing tends to arouse emotions more than the issue of performance. High Performance is, after all, the name of the game, so it is natural that every vendor wants to be able to tell a prospective customer: "My computer has the highest performance." But performance is not a single-dimensional phenomenon in HPC (or in just about anything else, for that matter). In addition to processor speed, memory size (and/or speed), input-output speed, and other factors usually have considerable bearing upon how a given computer system performs in a given situation. And if the situation changes, the relative performance of two different computer systems can also change -- drastically. Hence, using a single number to state performance capabilities is inherently inadequate and (therefore) misleading.

Nevertheless, users seem to dislike the complexity that goes with more accurate characterizations of capabilities, so the practice of using MIPS and megaflops persists. It is, of course, useful to know the "peak" megaflops rating of a supercomputer -- the maximum performance that could be attained if every element could be made to work in perfect harmony with every other element to bring the full theoretical capacity of the system to bear upon a problem -- even if it can never be attained in practice. It provides an upper bound on expectations, and it often sets a goal to be aimed at. Thus, a vendor can also brag about the efficiency of a particular software package in utilizing system resources.

But even this can be misleading, because what constitutes acceptable or superior efficiency in one situation may unacceptably inferior in another. This is especially true in comparing supercomputer systems which have very different architectures. For example, Japanese vector supercomputers have, until this year, always employed a single processor unit, whereas U.S. vector supercomputers have used as many as eight processors. However, the number of pipelines in Japanese processors have ranged up to 16, while U.S. systems almost invariably use two. As a result, the peak performance ratings of the most powerful Japanese vector supercomputers have been 2 to 16 times those of the most powerful single-processor U.S. vector supercomputers and up to double those of the largest multiple-processor U.S. systems.

In practice, however, the U.S. vector supercomputers tend to outperform their Japanese counterparts. For example, some recent benchmarks conducted on behalf of the U.S. Office of Naval Research (ONR) show that a single-processor Cray X-MP, rated at 235 peak megaflops, compares quite favorably with its contemporaries from Fujitsu and Hitachi, the VP-200 and S/810/20, even though the latter's peak performance more than twice as high: 500 megaflops and 630 megaflops, respectively (see Exhibit D-2). Likewise, the four-processor version of the X-MP, which is rated at 940 peak megaflops, is comparable to the (1,300 peak megaflops)NEC SX-2A and significantly faster than the newer Fujitsu VP-400E (1,700 peak megaflops). These benchmarks were based upon a widely-used technique for solving the Navier/Stokes equations (used in aerodynamics simulations), and the results were published in 1989 in a scientific periodical produced by ONR's Tokyo office.

Exhibit D-2: Actual vs. Peak Supercomputer Performance
(megaflops)



One reason for the discrepancy between peak and actual performance numbers in benchmarks such as this is that compilers may have difficulty in generating object code which uses the hardware with maximum efficiency. But even when special modifications to the algorithms are permitted to "tune" the benchmark programs to the architectures of the various machines (as was the case in the Navier/Stokes benchmarks cited above), the net ratios of actual-versus-peak performance still indicates that much of the potential capacity of the Japanese systems is going unused and may, indeed, be unusable for all practical purposes.

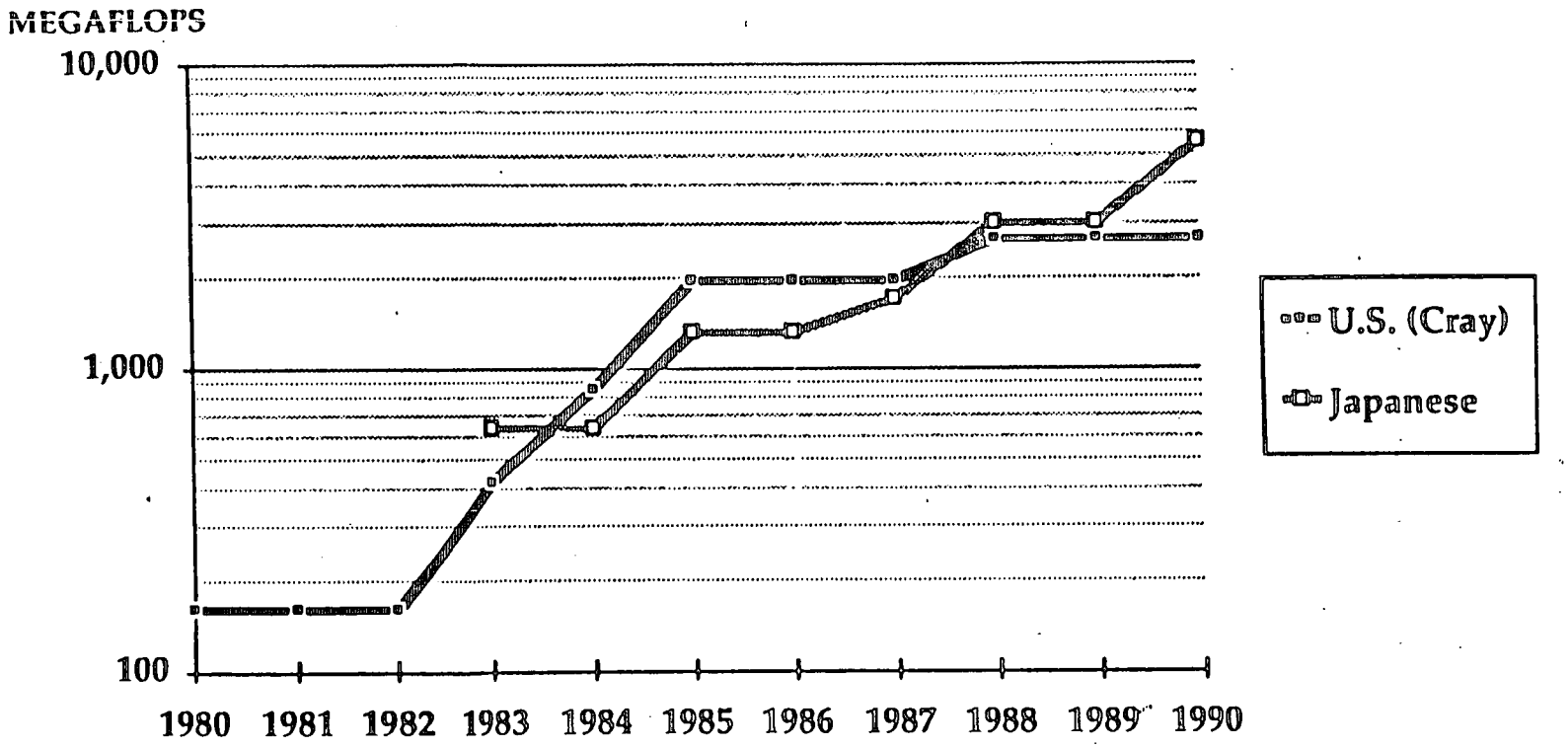
On the other hand, in multiprocessor systems such as the Cray X-MP, Y-MP, and Cray-2, "autotasking" compilers and operating systems must also work together to partition the computational load across all the processors in order to attain maximum peak performance, and as in the case of multiple pipes, it is impossible to attain 100 percent utilization in actual practice. Here the Japanese supercomputers have an inherent advantage over the Crays, because all of the Japanese systems, except the NEC SX-3, are single-processor systems. However, in the Navier/Stokes benchmark, an eight-processor Cray Y-MP achieved 84-88 percent utilization, which is quite respectable.

In general, it becomes increasingly difficult to efficiently utilize all available processing power as the number of pipelines increases and/or as the number of processors increases. (This can be stated quantitatively as Amdahl's Law.) This difficulty is one of the barriers to realizing the potential in highly-parallel supercomputers, especially massively-parallel systems. In specific instances, these systems have demonstrated very high performance levels: for example, a massively-parallel (65,536 processors) Connection Machine won the IEEE Computer Society's Gordon Bell prize for supercomputing performance in 1989 by achieving a sustained -- not peak -- performance level of 5.6 gigaflops in an application at Mobil Research and Development Laboratories. This is more than twice the peak speed of the fastest Cray currently available, the Y-MP8, but it is well below the peak performance potential, about 28 gigaflops, of the CM-2. However, when the criteria for the Gordon Bell prize were changed in 1990 to reflect actual performance over a range of 13 engineering and scientific problems, rather than in just a single problem, the Cray Y-MP8, with 2.67 peak gigaflops, once again emerged as the winner.

The foregoing discussion serves to illustrate an important point about High Performance Computing: the software and the algorithms used can have as great an impact upon "actual" performance as peak hardware speed, especially in the more complex architectures (involving greater degrees of parallelism) which are expected to dominate HPC in the future. (The recent experience of NASA's Langley Research Center provides another example. A computational model of the space shuttle solid rocket booster, the component that failed in the Challenger disaster, required 14 hours to run on a DEC VAX 11/780, a minicomputer that is widely used in research laboratories. In "unvectorized" form, it required one hour on a Cray-2, but after "vectorization" by Cray's compiler, it ran in just 14 minutes. Further optimization by a programmer brought the time down to 13 seconds, and parallelization on a Cray Y-MP cut it to just five seconds.) Hence, although the HPCC Program component which focuses upon advancing computer hardware is certainly essential, the potential payback from the software component is even greater in the longer term.

At present, the architectures of the large Fujitsu, Hitachi, and NEC supercomputers suggest that Japanese proficiency in multiprocessing still lags well behind that of the U.S. and that Cray is still the world leader in terms of overall system performance. (This does not include system "ease of use," which is another area in which the U.S. excels.) But the Japanese will learn, because there is no embargo on this know-how, and it is too late to impose one, even if such were desirable or feasible. At a minimum, we expect the Japanese gradually to move ahead of the U.S. in theoretical peak performance of vector supercomputers (see Exhibit D-3). The high degree of vertical integration of their companies -- Fujitsu, Hitachi, and NEC are also semiconductor manufacturers, NEC being the world's largest -- coupled with their national resolve, their history of extensive governmental support, and their intense domestic competition, will skew the odds significantly in their favor.

Exhibit D-3: Peak Supercomputer Performance

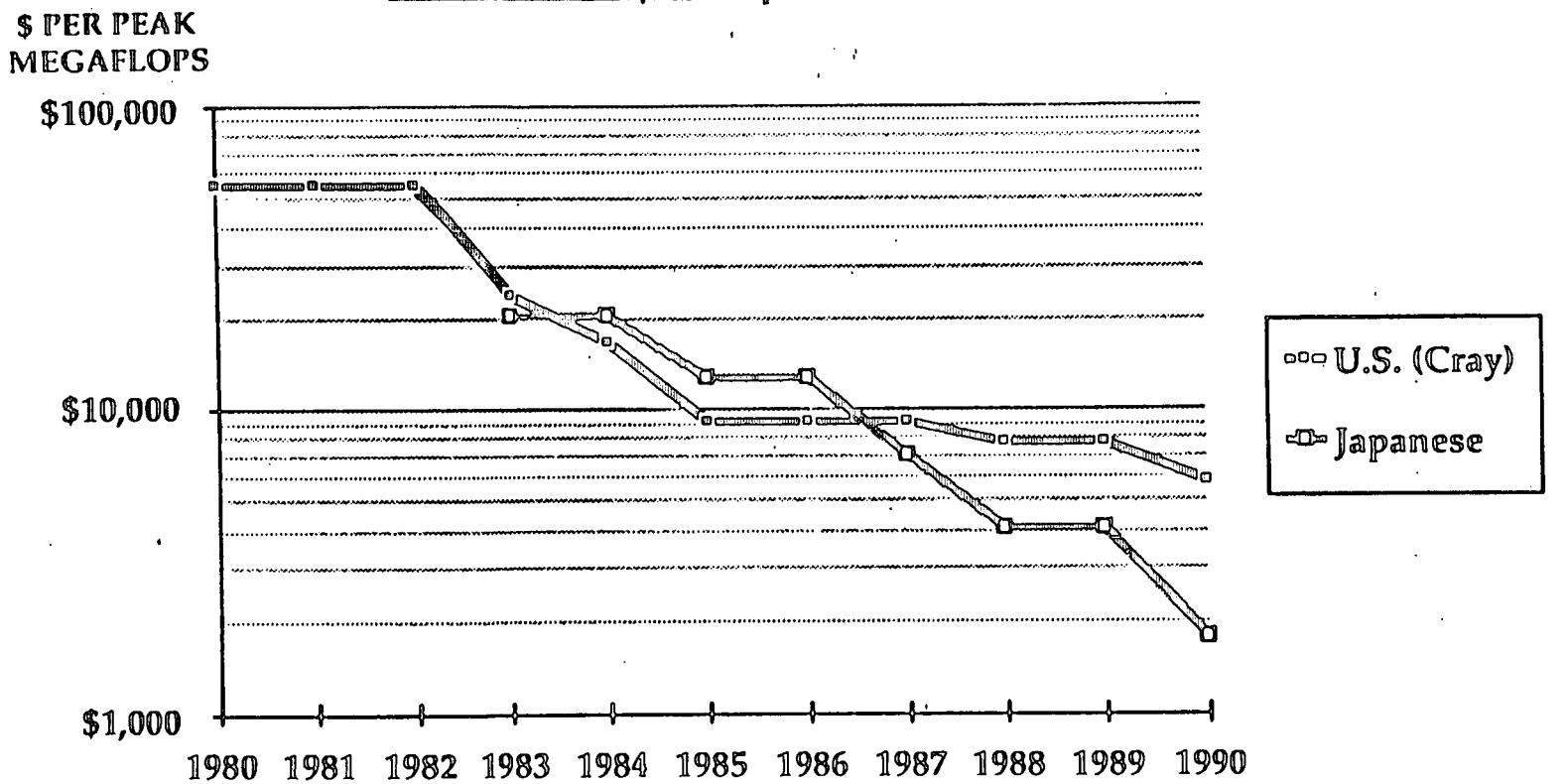


If peak hardware performance is not enough to carry the day, the Japanese may also point to price/performance* superiority as well (see Exhibit D-4). Although discounting practices in the Japanese market make comparisons between vendors well nigh impossible, the quoted prices for the U.S. market suggest that NEC's systems may offer as much as four times the price/performance of the Cray Y-MP. The Cray-3 and the Y-MP16 should close this gap somewhat, but the Japanese are expected to maintain a price/performance advantage over the U.S. in vector supercomputers for the foreseeable future.

Another competitive issue could be reliability, which has been running at more than 5,000 hours MTBF (Mean Time Between Failures) for the Japanese machines, as compared with less than 1,000 for the Cray X-MP series and worse than that for the Cray-2. Again, the Cray Y-MP series, which is VLSI based, should be markedly better than Cray's previous generations, but it is likely to remain behind the Japanese.

* Again, it must be emphasized that this comparison is based upon peak performance and, hence, may be very misleading. The best way to compare supercomputer systems is on the basis of how well they perform in the particular application(s) which the user wishes to do.

Exhibit D-4: Supercomputer Price/Performance



If there is a dark horse that may yet win the supercomputing race for the U.S., it is parallel systems. In terms of peak performance and price/(peak) performance, the systems now marketed by Intel and Thinking Machines are significantly better than the best vector supercomputers. Despite the formidable software problems associated with large-scale parallelism, Cray Research recently announced that it is embarking upon a development program in highly-parallel systems, and IBM is known to have some on-going development efforts in this area as well. Although Fujitsu announced in June, 1990, that it would begin shipping a 12.5 gigaflops, 1,024-processor parallel system, dubbed the CAP-II, in March, 1991, and other Japanese supercomputer makers are known to have similar prototype systems under development and/or in operation, America has a clear lead in this small, but rapidly growing, sub-segment of HPC at present.

Although it would be inappropriate for the Federal HPCC Program to emphasize the parallel approach to supercomputing at the expense of other more established approaches, the rapidity with which the Japanese overtook the U.S. in peak vector supercomputer performance in the past decade suggests that we must act expeditiously to avoid squandering the leads we still have: in parallel supercomputer architectures and in software for all types of supercomputers.

WORKSTATIONS AND NETWORKS

Until the late 1960s, supercomputer usage was "batch" mode (as was most other computing): users prepared programs in decks of punched cards, submitted them to the computer center dispatcher, and returned later to pick up the results. But with the advent of computer time-sharing, interactive usage, via a desktop workstation, became the preferred mode. Initially, this was limited to program development and debugging, with "production" runs still being in batch mode, often during night shifts, but today interactive operation is the norm. Not only does interactive usage offer the user the psychological benefits of near-instantaneous response, but it also allows the results of the computation to be viewed pictorially, rather than as page after page of numerical tables. Greatly improved processing and storage capabilities in workstations have even permitted full-color animation, which enables the user to gain intuitive insights into the processes or phenomena under investigation. The end result is significant enhancement of user creativity and productivity.

Of course, the workstation need not be located in the same room as the supercomputer. Since the advent of computer networks such as the ARPANET in the 1970s, users can gain access to supercomputer facilities halfway around the world, and they can share programs, data, and ideas with others as they do so. However, the capacity of available networks has not kept pace with demand in recent years, especially with the bandwidth requirements (estimated at 15 megabytes per second) of full-color workstations capable of animation. Although the National Science Foundation is now engaged in a major network upgrading program, many academic networks today are capable of no more than 200 kilobytes per second at the trunk level (the highest capacity portion of the network), and thus are totally inadequate for the kind of human-machine interaction needed. Thus, the lack of high-capacity networking facilities by which researchers can gain access to HPC centers is an even greater impediment to HPC usage today than the availability of supercomputers and processing time on them.

Exhibit D-5: A Thumbnail Sketch of HPC History

1940-50:	One-of-a-kind computers for WWII and after: Mark I, EDVAC, ENIAC, etc.
1950-54:	The first commercial HPC systems: ERA 1101 & 1103; IBM 701.
1959-64:	Limited-production commercial systems for U.S. defense laboratories: Univac LARC; IBM 7030 (Stretch); Control Data (CDC) 6600. LARC and Stretch had overlapped execution of multiple instructions; 6600 had multiple function units, instruction stack (cache).
1965-69:	Another generation of commercial systems for defense labs: CDC 7600; IBM 360/91.
1970-74:	Few-of-a-kind special systems: ILLIAC IV and Goodyear STARAN parallel array processors; Texas Instruments ASC and CDC STAR-100 multi-pipeline vector processors.
1975:	Floating Point Systems (FPS) begins production of commercial array processors to enhance performance of IBM, DEC, and other systems.
1976:	Commercial supercomputer production begins in earnest: first Cray-1 supercomputer installed.

... continued on next page

Exhibit D-5 (cont'd)

1981:	CDC enters the commercial supercomputer market: CYBER-205. Apollo begins selling technical workstations. FPS begins shipping "specialized scientific computers" (a.k.a. "minisupercomputers").
1983:	First multi-processor vector supercomputer installed: the Cray X-MP. First Japanese supercomputers installed: Hitachi S-810/20 and Fujitsu VP-200.
1984:	First 4-processor X-MP installed.
1985:	Convex and Alliant begin shipping minisupercomputers. First Cray-2 installed. First NEC supercomputer installed: over 1 gigaflops, single processor.
1986:	IBM begins installing Vector Facility on some 3090 systems First massively-parallel Connection Machine installed. Culler introduces the Personal Supercomputer: more power than a CDC 6600 for less than $100,000.
1987:	First installation of an ETA10: ultra-large-scale integration CMOS circuitry.
1988:	First Cray Y-MP installation: eight processors, parallelizing compiler.
1989:	The second generation of Japanese supercomputers begins: Hitachi S-820/80. CDC's subsidiary, ETA, drops out of the supercomputer business.
1991:	First Japanese multi-processor supercomputer, the NEC SX-3, to be installed. First *GaAs*-based systems to be shipped by Convex and Cray Computer.

[This page has been left blank intentionally.]

APPENDIX E - THE ROLE OF HPC

[This page has been left blank intentionally.]

THE ROLE OF HPC

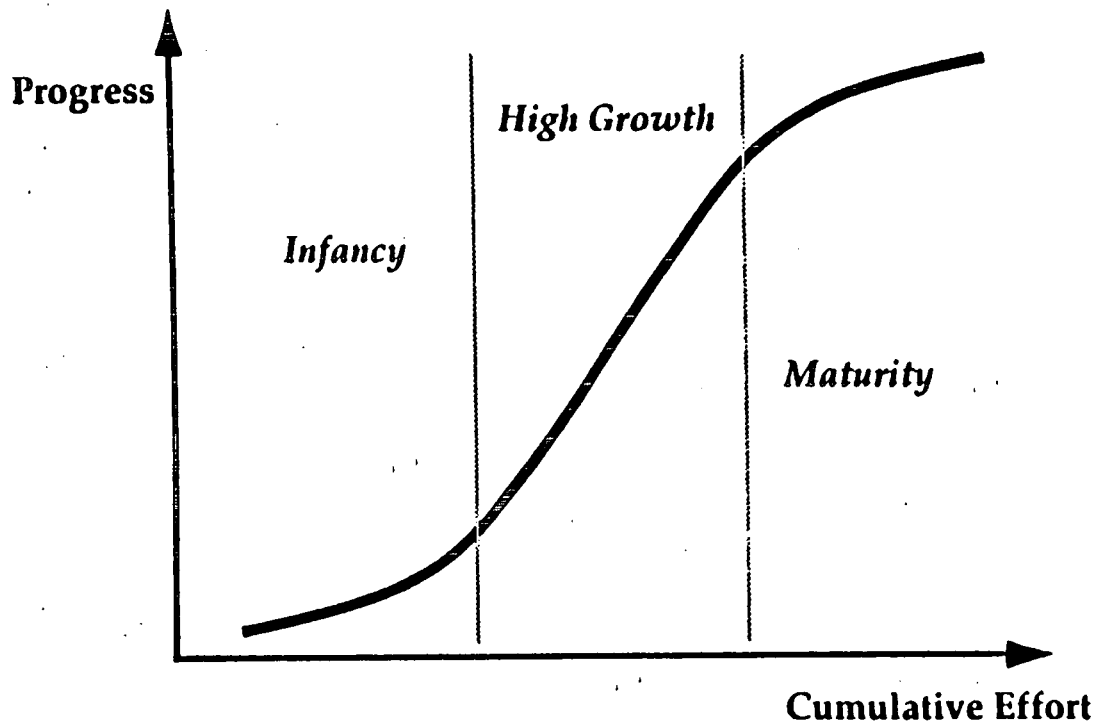
The history of the information industry can be characterized as one of incessant, often dramatic, change. Indeed, the industry has been likened to riding a bicycle: if it is not going forward rapidly enough, it will topple over. But information technology has also become a powerful agent of change in other industries as well, as witness the consequences of automation. Hence, there is a "Leverage Principle" at work here: a seemingly small change in information technology can have a significant effect upon the information industry, and that, in turn, can have an even larger effect upon those industries which use information systems. For example, the development of the microprocessor has led to desktop -- and will soon lead to handheld -- computing, which will revolutionize the ways in which companies conduct their business, how they are organized, and even what businesses they choose to engage in.

Where High Performance Computing (HPC) fits into all this follows from the fact that it is the part of the information industry where change is occurring most rapidly, the "cutting edge." (Indeed, this characterization could probably suffice to define HPC.) Hence, a change in HPC "compounds the leverage" mentioned above.

A CONCEPTUAL FRAMEWORK

Exhibit E-1 may help explain the role of HPC in advancing the state-of- the-art of computing in general:

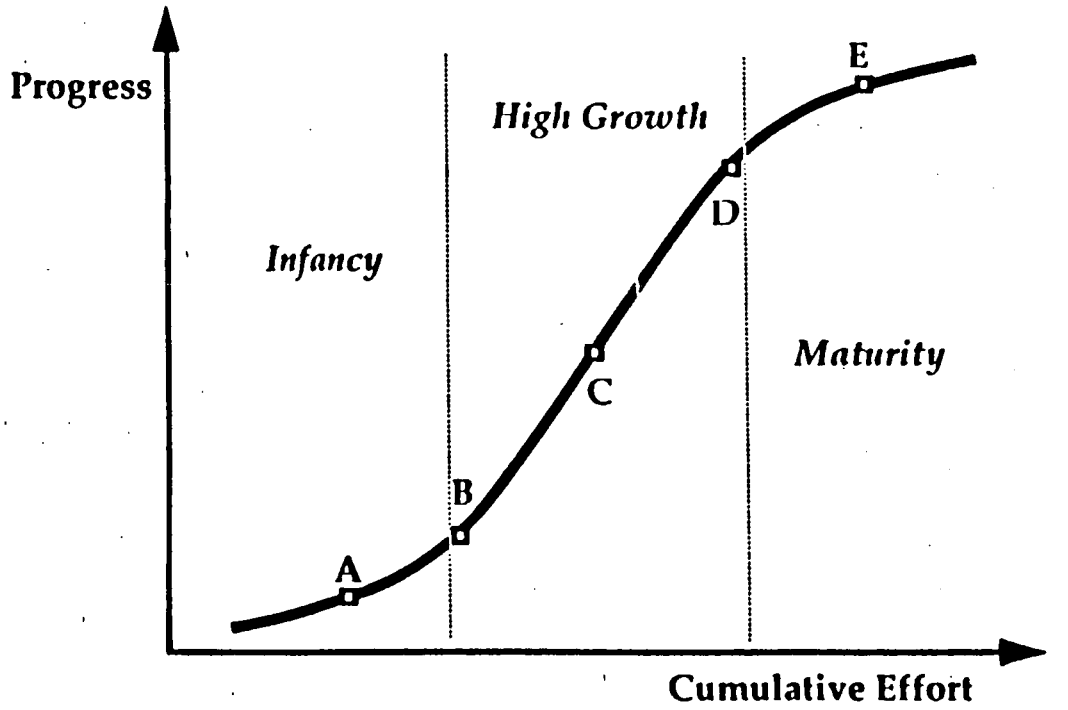
Exhibit E-1: Stages in Technological Development



The quintessence of High Performance Computing is to take advanced computing technologies from the "Infancy" stage and bring them into the "High Growth" stage.

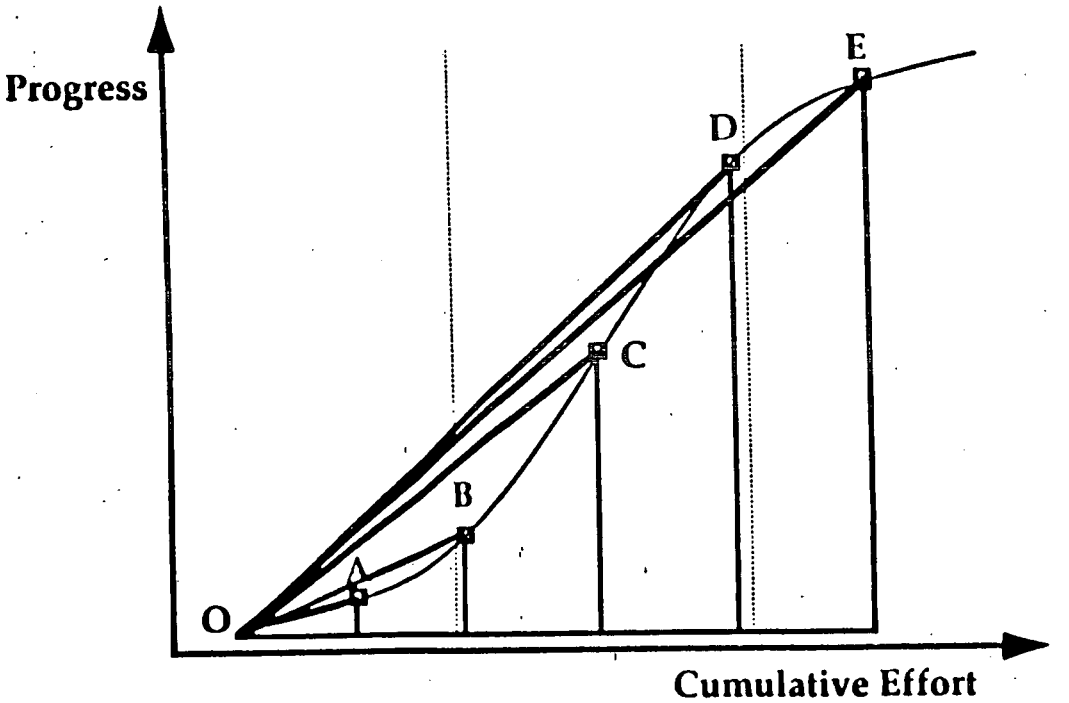
But there are problems associated with this. As shown in Exhibit E-1, technological advances require a disproportionally large amount of effort to bring about when the technology is in the "Infancy" stage or the "Maturity" stage, as compared with the intermediate "High Growth" stage. Hence, from a business standpoint, the optimal time to invest in a new technology is when it is at point "B" in Exhibit E-2.

Exhibit E-2: Timing of Technological Investment



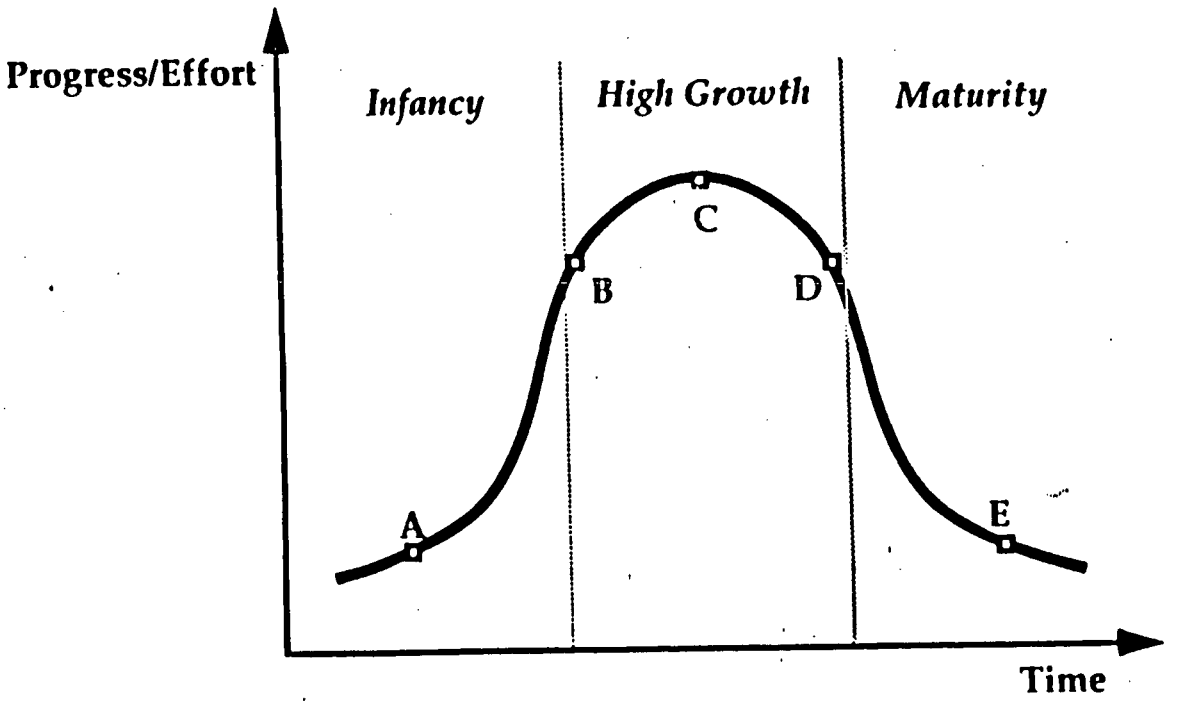
However, a technologically unsophisticated person, looking at historical Return on Investment (ROI), would tend to pass up "Infant" technologies and focus upon "High Growth" or even "Mature" ones because, based upon "track record," technologies at points "C", "D", and even "E" are more attractive: that is, the lines from "O" to "C", "D", and "E" have steeper slopes than the line from "O" to "B" in Exhibit E-3. The flaw in this thinking is, of course, unwarranted extrapolation: assuming that the future will be the same as the past.

Exhibit E-3: Return on Technological Investment



On the other hand, if the technological "progress"curve is redrawn as Exhibit E-4, then it becomes clear where the optimal return on investment lies.

Exhibit E-4; Optimal Return on Technological Investment



However, getting business to invest at point "A" is very difficult. (This is the "pre-competitive" stage.) Because of the low payoff-per-effort-invested, government incentives and support may be required if anything significant is to happen in this area. But once a technology reaches point "B," it is time for private industry to take over and for government to get out of the picture, except as necessary to transfer the technology to the private sector as expeditiously as possible. This is the model used by the Japanese government in its computer-related (and other) R&D initiatives since World War II. Whatever level of "success" may be ascribed to these various projects in attaining their ostensible goals, it is undeniable that the resultant transfer of technology into the Japanese computer industry and also into some key user industries -- what the U.S. Department of Defense would call "technology insertion" -- has been very, very efficient and effective. That fact alone may be sufficient to explain Japan's rapid ascendancy in the worldwide computing milieu and perhaps much of its growth as an industrial power. (Three of the world's top six computing firms are now Japanese whereas there were no Japanese companies in the "top 20" in 1970.)

The corollary is this: If U.S. firms are forced to get their key technologies from foreign sources, the technology transfer may not be as efficient as it is to their foreign competitors. For example, our firms might get the technologies at points "C" or "D" instead of point "B," where their foreign competitors would be able to buy in. Or in an extreme case, foreign governments might withhold technologies until they reach point "D," just as U.S. export controls have attempted to do with strategic technologies during the Cold War.* The implications for competitiveness are obvious: foreign firms would be able to bring products embodying new and superior technologies to market earlier and thereby capture greater market shares, and they would be able to spread their technological expenditures over longer product cycles, thereby enabling lower prices and/or greater profits.

---

* Alternatively, foreign suppliers might engage in a two-tier pricing strategy, charging U.S. customers more for their technologies (in contrast to recent alleged "dumping" practices) and thereby putting our firms at a cost disadvantage, or foreign governments might simply use the threat of any of these actions as a major bargaining point in trade negotiations with the U.S.

HPC "LEVERAGE"

The consequences of the foregoing discussion reach well beyond just the computer industry. Because of the obvious importance of computer systems in computational science and the growing importance of computational science in bringing new products to market in a number of industrial sectors, any weakness in HPC becomes amplified throughout a nation's industry: science does not advance at full speed; new products are more expensive to develop and take longer to bring to market; new concepts and ideas go unexplored. Thus, a small change in HPC investment can, via the "Trickle-Down Effect," make a significant difference in a broad range of computer systems applications and in the level of computational science, and this, in turn, can directly and profoundly affect national competitiveness, productivity, quality of life, etc.

[This page has been left blank intentionally.]

APPENDIX F - THE ECONOMICS OF HPC

[This page has been left blank intentionally.]

THE ECONOMICS OF HPC

The economics of HPC are poorly understood, even among those who are established supercomputer users, but especially among those who are not. Supercomputing is believed to be expensive, which probably has its basis in the relatively high initial cost of supercomputers. The largest Cray Y-MP models are priced at slightly over $20.5 million, which is enough to buy more than 1,400 of the most powerful single-user workstations available today, IBM's RS/6000 Model 320. The power of just one of these workstations is 40 peak megaflops, so the combined power of 1,400 of them would be 56,000 peak megaflops, or about 21 times that of the largest Cray. Based upon this performance comparison, plus the difficulties inherent in serving 1,400 users with a single Cray, most executives would opt to purchase the workstations, not the Cray. But this could be the wrong decision.

Implicit in the foregoing analysis is the assumption that an RS/6000 workstation will be powerful enough to solve any problem which arises. Although it is true that 40 megaflops is roughly equivalent to the most powerful supercomputers available in the mid-1960s, there are obviously some problems which require more power than this. Unfortunately, it is not yet practicable to link together several workstations to bring more processing capacity to bear (although early versions of software systems intended to that are just now becoming available), so the comparison which must be made is between the problems which can be solved on a single workstation versus those which can be solved on a current-generation supercomputer.

EXAMPLE #1: POSSIBILITY

Let us suppose that a researcher has two months (40 working days) to solve a particular problem, working "iteratively" with the aid of a supercomputer. If it is a "typical" supercomputer-class problem, the researcher's regimen might be something like this:

(a) Work all day on the problem, preparing a supercomputer job to be processed overnight.

(b) Have the job run during the night shift, collect the results the next (workday) morning, and use them in step (a) the following day.

If we assume the researcher's time costs $100 per hour (including overhead) and Cray Y-MP8 time costs $1,200 per hour (an intentionally high estimate), the total budget for the project would be $512,000 (based upon 8 hours per day for the scientist and 10 hours per day for the Cray). The total amount of computing involved would be 3.84 billion megaflops. (Actually, it would be something less than this, because it is impossible to utilize the full theoretical "peak" capacity of any computer. See Appendix D.)

Now let us suppose that an RS/6000 is substituted for the Cray. If the total computing provided by the Cray is divided by the number of days (40), the amount of computing capacity required per researcher-workday is 96 million (peak) megaflops. This is equivalent to more than 665 hours of processing on an RS/6000 Model 320, so the researcher's regimen would become the following:

(a) Work all day on the problem, preparing an RS/6000 job to be started at the end of the day.

(b) Let the RS/6000 run overnight and all day, 24 hours per day, for 28 more days. Then take the results and return to step (a).

To complete forty such "cycles" under these circumstances would require 1,160 days, or almost 3.2 years. And while it is true that the computing cost and the total budget would have been cut by more than $450,000 through substituting the RS/6000 for the Cray -- the former's cost is computed at $0.83 per hour -- this is clearly not a viable way to solve a problem of this magnitude.

This explains why current generations of "mainstream" computer systems, which may exceed the performance of earlier generations of supercomputers, cannot be substituted for modern high performance computers. The latter are at the cutting edge of performance, usually employing the most advanced technologies, and hence the capabilities they provide are not just quantitatively, but qualitatively, different from current run-of-the-mill systems. To put this another way, an army of mathematicians, each equipped with an electronic calculator, could in theory provide the same number of megaflops as a modern-day supercomputer system, but nobody would consider using the human horde approach to carry out the calculations necessary to simulate an automobile crashing into a wall! Experience has shown that scientists and engineers will not even attempt to solve problems which they know will require more than a few weeks or months of computing time, and even then only in extreme cases; most computer users want their results in seconds or minutes. Hence, HPC allows users to attack and solve problems which they wouldn't even consider otherwise, and thereby contributes significantly to advancing the state of the art in any industry or discipline where it is applied.

But, of course, not all problems are of this magnitude. Indeed, only a small portion of them are. So let us now run the foregoing analysis in reverse by considering a "more realistic" problem, based on "typical" workstation usage.

EXAMPLE #2: TIME

As before, we assume that a researcher has two months (40 working days) to solve a problem, working "iteratively" with an RS/6000. The researcher's daily regimen might be something like this:

(a) Work one hour on the problem, preparing an RS/6000 job to be started at the end of the day.

(b) Let the RS/6000 run overnight, collect the results the next (workday) morning, and use them in step (a) the following day.

The total cost of the project, assuming the same rates for as before, would be $4,332. The total amount of computing (based upon 10 hours per night) would be 57.6 million (peak) megaflops.

However, if a Cray were substituted for the RS/6000, the computing that was done overnight (1,440,000 peak megaflops) could be done in 9 minutes. Hence, the researcher's think-compute regimen could be collapsed, and as many as seven "cycles" could be completed in a single workday. At that rate, the entire project would be completed in just over 1 week instead of two months.

In a competitive environment, this could be critical in beating the competition. Indeed, as explained by Professor Robert M. Hayes of the Harvard Business School:

EXAMPLE #2: TIME

"Product cost and quality (whether defined in terms of defect rates, tolerances, reliability, or lifetime cost of operation) are no longer as dominant in determining a company's success as they used to be. One must be careful not to fight the previous war -- as Marshall Foch warned France prior to World War II. Even as companies have gained rough parity with one another on these familiar dimensions, another has arisen to define a new competitive arena: speed. The speed with which existing products can be delivered, the speed with which customized versions of current products can be produced, the speed with which entirely new products can be created -- these are the new weapons of international competition....

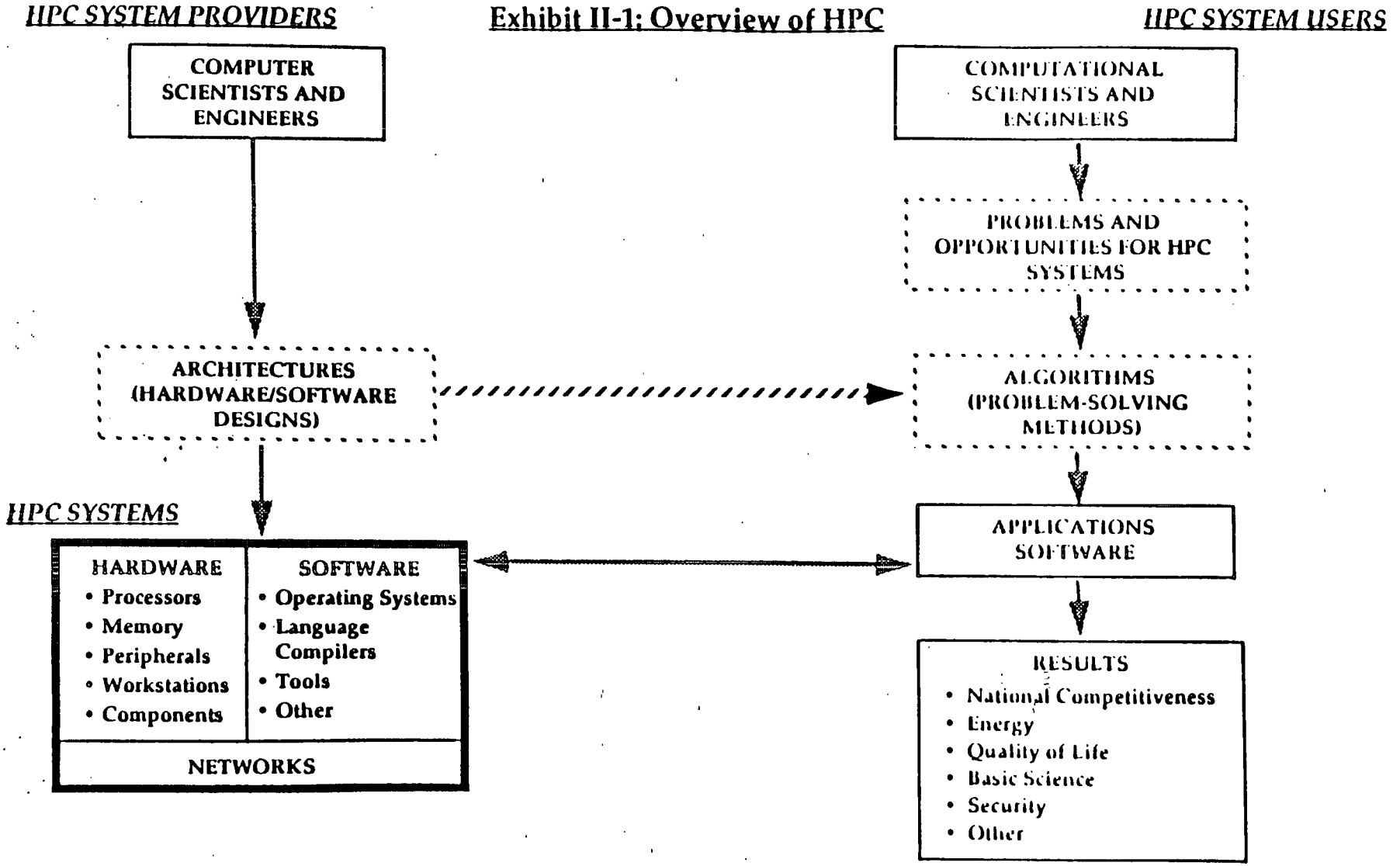
A company that can design and introduce a new product in half the time its competitors take can bring that product to market well ahead of them. It also has the possibility of waiting much longer, until technological choices and customer preferences have clarified, before beginning to design its next product. If it chooses the first option, within a couple of design cycles the faster company will be introducing new products that are a whole generation ahead of its competitors. The achievement of parity in manufacturing cost and quality, no matter how much time and effort was expended on that goal, has little value if the company can offer only obsolete products to its customers....

The same technologies that help designers and engineers shorten development and manufacturing time have also changed the character of product palettes. Not only can companies develop new products or modify existing ones much faster, but suddenly they can broaden their product lines without losing control of cost. This capability has become increasingly important -- even imperative -- in the last decade, as more and more companies attempt to elbow their way into markets that are already approaching saturation. Capitalizing on the fragmentation of consumer preference in our post-industrial society, many of these new competitors have sought toe-holds by designing and manufacturing variations of current products. Unless existing companies match these product offerings, they risk losing sales to the new entrants. Companies are no longer content simply to look for market niches where they will be relatively free of competitive pressure; today they seek out 'micro-niches.' The result is that most companies today find themselves offering a greater variety of products, each at lower volume, than they used to....

Source: **Designing for Product Success;** The Design Management Institute, 1990.

SUMMARY: BACKGROUND (Chapter III)

Exhibit II-1: Overview of HPC



SUMMARY

These three examples are summarized in Exhibit F-1. The "bottom line" resulting from the analysis is clear:

Supercomputing may be more expensive than workstation-based computing in terms of actual cash outlay, but when viewed in the context of competitive situations, it is often the best (if not the only) choice.

Moreover, when viewed from an enterprise-wide perspective, the significance of supercomputing is even greater. For each researcher in the above examples, a typical company might have 10, or 100, or even 1000 other employees -- production workers, sales and support people, etc. -- whose jobs are dependent upon the solution(s) obtained. Hence, the ability to obtain a solution at all (Example #1), or to obtain it faster (Example #2), or to obtain a better solution (Example #3) can have a very broad impact, throughout the company. This is the "leverage" that makes HPC a critical element in industrial competitiveness.

Exhibit F-1: Supercomputers vs. Workstations, Cost vs. Effectiveness of Solutions

		Cost	Megaflops	Time
Example #1 (Possibility of Solution)	IBM RS/6000 Cray Y-MP8 Cray : IBM	$54,136 $512,000 9.46 : 1	3,840,000,000 3,840,000,000 1 : 1	3.18 years 40 working days 1 : 29
Example #2 (Time of Solution)	IBM RS/6000 Cray Y-MP8 Cray : IBM	$4,332 $11,199 2.59 : 1	57,600,000 57,600,000 1 : 1	40 working days 6 working days 1 : 6.67
Example #3 (Quality of Solution)	IBM RS/6000 Cray Y-MP8 Cray : IBM	$32,332 $80,000 2.47 : 1	57,600,000 384,048,000 6.67 : 1	40 working days 40 working days 1 : 1

THE PARALLEL POTENTIAL

But there is the potential for even greater advantage, and it lies in the emerging form of supercomputing based upon highly-parallel systems. Perhaps the most successful (so far) representative of this genre is the CM-2 "Connection Machine," developed and made by Thinking Machines Corporation. The price of the largest configuration (65,536 processors) is about $8.5 million, or just over 40 percent of a Cray Y-MP8, but its "peak" processing power is more than 10 times greater, in excess of 28,000 megaflops. On a price/peak performance basis, the CM-2 is more than 25 times better than the Cray, and it is even 15 percent better than the RS/6000.

The problem is that, because of its radically different -- that is, massively-parallel -- architecture, the CM-2 is difficult to use on many classes of problems where the Cray and the RS/6000 are established performers. In a number of instances, however, the applicability of the CM-2 has been demonstrated, and the results have been impressive. Indeed, TMC and one of its customers, Mobil Research and Development Corporation, won the IEEE Computer Society's Gordon Bell prize for supercomputing performance in 1989 by achieving a sustained performance level of 5.6 gigaflops (1 gigaflops = 1,000 megaflops) in a seismic processing application. This is well below the peak performance potential, about 28 Gigaflops, of the 65,536-processor Connection Machine, but it is more than double the peak speed of the fastest Cray currently available.

An important challenge now before the HPC community is to find ways to use the CM-2 (and other systems employing parallelism) more efficiently and in a wider range of problems*. Ultimately, this may entail variants of parallel processing that are quite removed from the CM-2 architecture, perhaps even closer to the Cray's "vector pipeline" design, but the point is that there is a major opportunity ahead in HPC. What is needed is an initiative, such as the Federal HPCC Program is intended to provide, to take advantage of this opportunity.

* This may help users of "traditional" vector supercomputers as well. Experience has shown that, when applications programs are adapted for massively parallel systems, the resulting "modified" programs also run significantly faster on vector (and even scalar) systems than the "original" versions.

To help quantify the magnitude of the potential benefits, let us return to the three examples given above, this time substituting a CM-2 for the Cray, as if that could be done in a simple and straightforward fashion (which it presently cannot in many applications because of software problems).

Example #1 ("Possibility of Solution"): The CM-2 could provide the equivalent of 10 hours/day of Cray processing in about 0.94 hours (56+ minutes) per day. At $500 per hour, the computing cost would be cut drastically, bringing overall project costs (for the same amount of "work") to less than 10 percent of the previous (Cray-based) level. Alternatively, the CM-2 could be run 10 hours per night like the Cray, providing more than 10 times the total (peak) computing, but with overall project costs reduced by more than half.

Example #2 ("Time of Solution"): The CM-2 would perform the 1,440,000 (peak) megaflops per day in less than 1 minute, allowing shortening of total project time as with the Cray (perhaps even more so, if the researcher were willing to work a few minutes overtime each day). The total cost would be slightly less than with the RS/6000.

Example #3 ("Quality of Solution"): The CM-2 would provide in 1 hour per day more than 10 times the (peak) computing of the Cray in the same amount of time, and the project cost would be 35 percent lower. Alternatively, it would provide in 1 minute per day the same amount of (peak) processing as the RS/6000 in 10 hours per day, but at slightly lower total project cost; and it would provide the same processing in under 6 minutes per day as the Cray would in 1 hour per day, but at less than half the cost.

The following exhibit summarizes how the CM-2 compares with the Cray Y-MP8 and the IBM RS/6000 in these three examples.

Exhibit F-2: The Potential Advantages of Highly-Parallel Systems

		Cost	Megaflops	Time
Example #1 (Possibility of Solution)	IBM RS/6000 Cray Y-MP8 TMC CM-2	$54,136 $512,000 $50,716	3,840,000,000 3,840,000,000 3,840,000,000	3.18 years 40 working days 40 working days
Example #2 ("Time of Solution)	IBM RS/6000 Cray Y-MP8 TMC CM-2	$4,332 $11,199 $4,281	57,600,000 57,600,000 57,600,000	40 working days 6 working days ≤6 working days
Example #3 (Quality of Solution)	IBM RS/6000 Cray Y-MP8 TMC CM-2	$32,332 $80,000 $52,000	57,600,000 384,048,000 4,100,000,000	40 working days 40 working days 40 working days

As mentioned before, it is not, in general, feasible to substitute a CM-2 (or any other highly-parallel supercomputer) for a Y-MP (or any other vector supercomputer) at today's state-of-the-art. But the feasibility has been demonstrated in a number of applications, and that number is growing everyday. Even in these applications, the "speed-up" of the parallel over the vector system may not be as great as their relative peak megaflops ratings would suggest, but it is substantial nonetheless.

Thus, highly-parallel supercomputers are seen as having the potential to provide unprecedented levels of processing power, at drastically improved price/performance levels. The challenge, which the Federal HPCC Program is intended in part to address, is to realize this potential and to keep the U.S. ahead of foreign competitors in this increasingly important segment of HPC.

[This page has been left blank intentionally.]

APPENDIX G - APPLICATION OPPORTUNITIES

[This page has been left blank intentionally.]

Exhibit G-1: Application Opportunities

Legend: ● Primary Area ○ Secondary area

APPLICATION	DESCRIPTION	Category: COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	COMMENTS
• Materials Science	High performance computing has provided invaluable assistance in improving our understanding of the atomic nature of materials. These have an enormous impact on our national economy. A selected list of such materials includes: semiconductors, such as silicon and gallium arsenide, and superconductors such as the high Tc copper oxide ceramics that have been shown recently to conduct electricity at about 100 degrees Kelvin.	●			○	●		
• Semiconductor Design	As intrinsically faster materials such as gallium arsenide are used, a fundamental understanding is required of how they operate and how to change their characteristics. Essential understanding of overlay formation, trapped structural defects, and the effect of lattice mismatch on properties are needed. However, materials with defects and mixed atomic constituents are beyond present capabilities.	●			○	●		

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Category						
		COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	
• Turbulence	Turbulence in fluid flows impacts the stability and control, thermal characteristics, and fuel performance of virtually all aerospace vehicles. Understanding the fundamental physics of turbulence is requisite to reliably modeling flow turbulence for the analysis of realistic vehicle configuration.	○					●	Principally a NASA concern, but with wider applicability.
• Superconductivity	The discovery of high temperature superconductivity in 1986 has provided the potential for spectacular energy-efficient power transmission technologies, ultra-sensitive instrumentation, and devices using phenomena unique to superconductivity. The materials supporting high temperature superconductivity are difficult to form, stabilize, and use, and the basic properties of the superconductor must be elucidated through a vigorous fundamental research program.	○	○		●	○		This is one of the newest scientific frontiers.

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Category						
		COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	
• Efficiency of Combustion Systems	To attain significant improvements in combustion efficiencies requires understanding the interplay between the flows of the various substances involved and the quantum chemistry which causes those substances to react. In some complicated cases, the quantum chemistry required to understand the reactions is beyond the reach of current supercomputers.	○	●	○		○		Efficient use of dwindling hydrocarbon fuels is becoming crucial.
• Enhanced Oil and Gas Recovery	This challenge has two parts: to locate as much of the estimated 300 billion barrels of oil reserves in the U.S. as possible and then to devise economic ways of extracting as much of this as possible. Improved seismic analysis techniques as well as improved understanding of fluid flow through geological structures is required.	○	●	○		○		

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Competitiveness	Energy	Quality of Life	Basic Science	National Security	Other	
• Nuclear Fusion	Development of controlled nuclear fusion requires understanding the behavior of fully ionized gasses at very high temperatures under the influence of strong magnetic fields in complex three dimensional geometries.	○	●	○		○		A long-term effort.
• Design of Pharmaceuticals	Predictions of the folded conformation of proteins and of RNA molecules by computer simulation is rapidly becoming accepted as a useful, and sometimes primary, tool in understanding the properties required in pharmaceutical design.	○		●				Eli Lilly recently became the first U.S. pharmaceutical firm to acquire its own supercomputer.

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Category						
		COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	
• Structural Biology	This function of biologically important molecules can be simulated by computationally intensive Monte Carlo methods in combination with NMR of crystallographic data. Molecular dynamics methods are required for the time dependent behavior of such macromolecules. The determination, visualization, and analysis of these 3D structures is essential to the understanding of the mechanisms of enzymic catalysts, recognition of nucleic acids by proteins, antibody/antigen binding, and many other dynamic events central to cell biology.				○	●		
• Human Genome	Comparison of normal and pathological molecular sequences is our current most revealing computational method for understanding genomes and the molecular basis for disease. To benefit from the entire sequence of a single human will require capabilities for more than three billion subgenomic units, as contrasted with the ten to two hundred thousand units of typical viruses.				●	○		Requires supercomputing power far beyond today's capabilities. Potential to find cures for major diseases.

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Competitiveness	Energy	Quality of Life	Basic Science	National Security	Other	
• Prediction of Weather, Climate and Global Change	The aim is to understand the coupled atmosphere, ocean, biosphere system in enough detail to be able to make long-range predictions about its behavior. Applications include: understanding CO dynamics in the atmosphere, ozone depletion, climatological perturbations due to man-made releases of chemicals or energy into one of the component systems, and detailed predictions of conditions in support of military missions.			●	○	○		A truly "global" application with potentially far-reaching implications.
• Computational Ocean Sciences	The objective is to develop a global ocean prediction model incorporating temperature, chemical composition, circulation, and coupling to the atmosphere and other oceanographic features. This will couple to models of the atmosphere in the effort on global weather as well as having specific implications for physical oceanography.			○	●			Closely related to global climate prediction.

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Category						
		COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	
• Astronomy	Data volumes generated by Very Large Array (VLA) or Very Long Baseline Array (VLBA) radio telescopes currently overwhelm available computational resources. Greater computational power will significantly enhance their usefulness in exploring important problems in radio astronomy, resulting in a better return on a major national investment.				●			
• Quantum Chromodynamics	In high energy theoretical physics, computer simulations of QCD are yielding first-principle calculations of the properties of strongly interacting elementary particles. New phenomena have been predicted, including the existence of a new phase of matter, and the quark-gluon plasma. Properties under the conditions of the first microsecond of the Big Bang and in the cores of the largest stars have been calculated by simulation methods. Beyond the range of present experimental capabilities, computer simulations of grand unified "theories of everything" have been devised using QCD (Lattice Gauge Theory).				●			Requires speeds far beyond today's supercomputers.

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Category						
		COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	
• Speech	Speech research is aimed at providing a communications interface with computers based on spoken language. Automatic speech understanding by computer is a large modeling and search problem in which billions of computations are required to evaluate the many possibilities of what a person might have said within a particular context	●			○		○	Has potentially very broad applicability in the computer age. May also be useful in certain weapons systems.
• Vision	The challenge is to develop human-level visual capabilities for computers and robots. Machine vision requires image signal processing and reasoning. A competent vision system will likely involve the integration of all of these processes with close coupling.	●					○	Broad applicability in manufacturing, logistics processes and possibly weapons systems. Tougher problem than speech.

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Category						
		COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	
• Vehicle Signature	Reduction of vehicle signature (acoustic, electromagnetic, and thermal characteristics) is critical for low detection military vehicles.						●	
• Undersea Surveillance	The Navy faces a severe problem in maintaining a viable anti-submarine warfare (ASW) capability in the face of quantum improvements in Soviet submarine technology, which are projected to be so substantial that evolutionary improvements in detection systems will not restore sufficient capability to counter their advantages. An attractive solution to this problem involves revolutionary improvements in long-range undersea surveillance which will be computationally intensive.						●	

... continued on next page

Exhibit G-1 (cont'd)

Legend: ● Primary Area ○ Secondary area

		Category: COMPETITIVENESS	ENERGY	QUALITY OF LIFE	BASIC SCIENCE	NATIONAL SECURITY	OTHER	
• Engineering Applications	Structural analysis of products, including crash simulation of vehicles, fluid dynamics modeling of products and processes, etc. are important in reducing cost, achieving higher quality, and reducing time to market.	●	○			○		Aerospace, automotive and chemical industries are early adopters, but there is broad applicability across all industries.
• Computational Chemistry	Molecular modeling and simulation of chemical reactions underlie the development of new materials, electronics, pharmaceuticals, etc.	●	○	○	○			
• Film Animation	A perennial low-profile user of supercomputing power, this area will become more prominent with the shift to image-based computer applications in the 1990s.			●			○	
• Bond Bidding	Financial applications of supercomputers, based upon complex econometric models, should emerge in the 1990s.	●		○			○	Three Japanese securities firms have purchased supercomputers in the past two years.

[This page has been left blank intentionally.]

APPENDIX H - HPC HUMAN RESOURCES

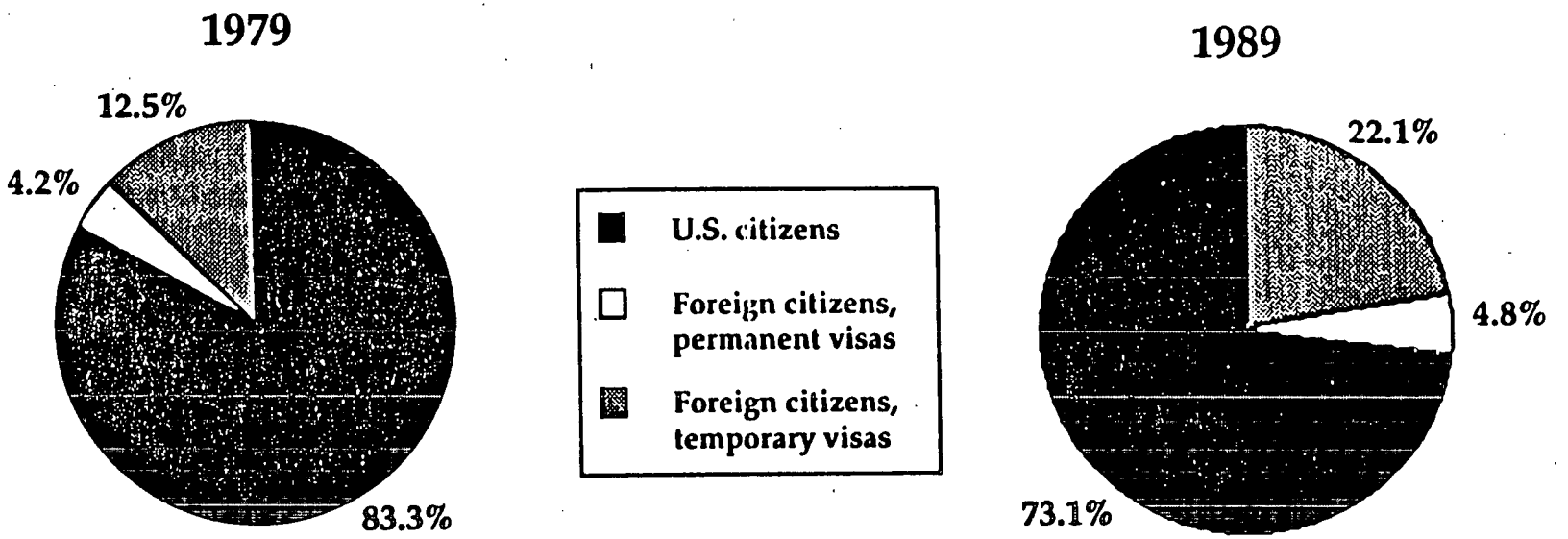
[This page has been left blank intentionally.]

HPC HUMAN RESOURCES

Notwithstanding the concerns with the various aspects of U.S. HPC leadership which are described above, there is one problem which tends to overshadow all others: the lack of adequately trained manpower. In recent years, the number of baccalaureate degrees in science and engineering awarded to U.S. citizens has leveled off or declined because of a decreasing college-age population. And since the mid-1960s, the rate at which students with natural science and engineering baccalaureate degrees from U.S. institutions went on to earn Ph.D.s has dropped by half. This reduction has been especially apparent among U.S. males, a group that has historically been the mainstay for doctoral degrees.

The recent growth in Ph.D. awards in several fields is largely due to greater participation by foreign students. In engineering, almost 60 percent of all doctorates are now awarded to foreign nationals, as are over a third of the doctorates in mathematics and physics. Approximately half of all foreign graduate students remain in the United States after getting their degrees, making valuable contributions to the nation's economy, research and education, but the likelihood that they will return to their homelands in increasing numbers to take advantage of improved career opportunities there raises serious questions about their continued availability.

Exhibit H-1: Science Doctorates Received in the U.S., 1979-1989



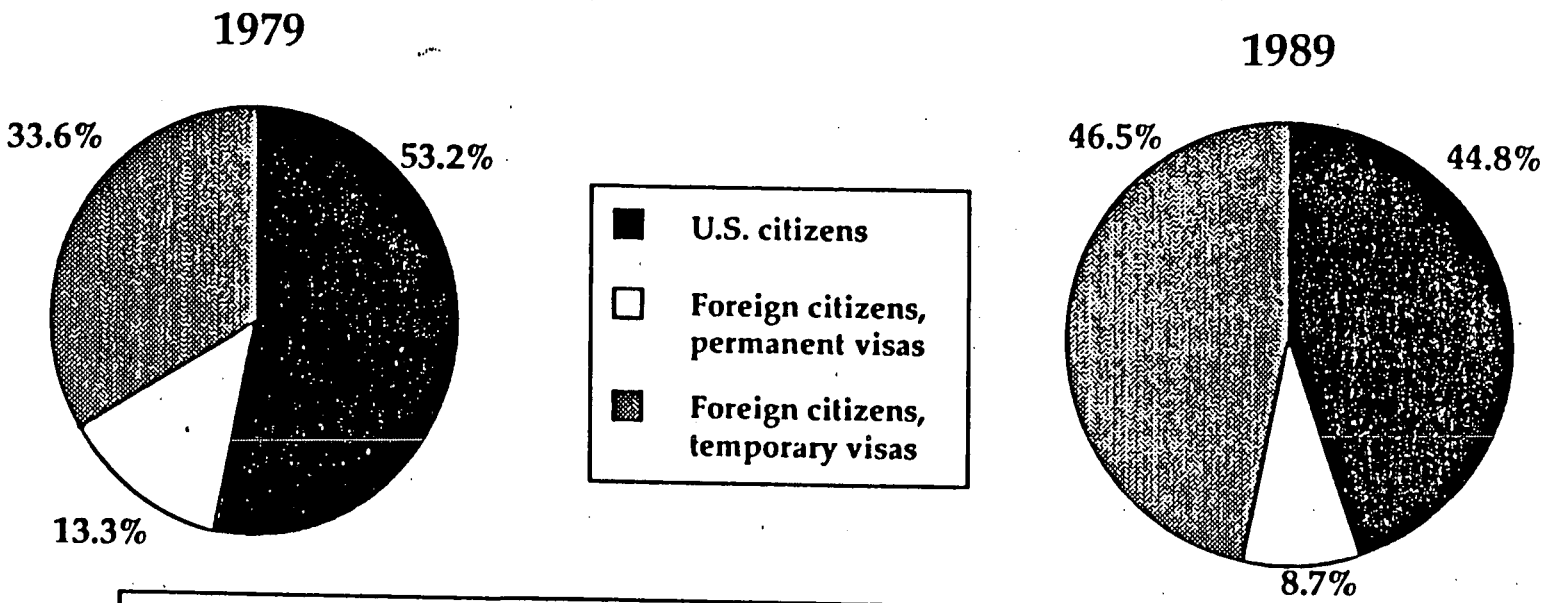
Note: Total Science includes: Physical Sciences; Earth, Atmospheric and Marine Sciences; Mathematics; Computer/ Information Sciences; Agricultural and Biological Sciences; Social Sciences; and Psychology.

Source: March, 1990, report of the Science and Engineering Education Sector Studies Group, National Science Foundation.

Exhibit H-1 shows the percentage distribution of U.S. science doctoral degrees awarded to U.S. students and non-U.S. students. The United States has gone from an 83.3 percent share in 1979 to 73.1 percent in 1989. Doctoral degrees to foreign students holding temporary visas, on the other hand, have increased from a 1979 share of 12.5 percent to 22.1 percent of doctoral degrees awarded in 1989. This represents a compound annual growth rate (CAGR) of 5.9 percent for holders of a temporary visas, and a negative 1.3 percent CAGR for U.S. students.

The situation in engineering is even worse. As Exhibit H-2 demonstrates, during the past ten years the share of engineering degrees obtained by holders of foreign temporary visas has grown from 34 percent in 1979 to 47 percent in 1989. The United States share has decreased from 53 percent in 1979 to 45 percent in 1989. The portion of engineering degree recipients with permanent visas fell from 13.3 percent in 1979 to 8.7 percent in 1989. In terms of nationality, Koreans have a CAGR of 23 percent over the 1979-1989 time frame, and Taiwan follows closely with a 16 percent CAGR, while India has gone from 142 doctoral degrees in 1979 to 216 doctoral degrees in 1989 (a CAGR of 4.3 percent). Japanese students, surprising as it seems, have dwindled from 26 recipients in 1979 to 19 doctorates granted in 1989 (a CAGR of negative 3.2 percent).

Exhibit H-2: Engineering Doctorates Received in the U.S., 1979-1989

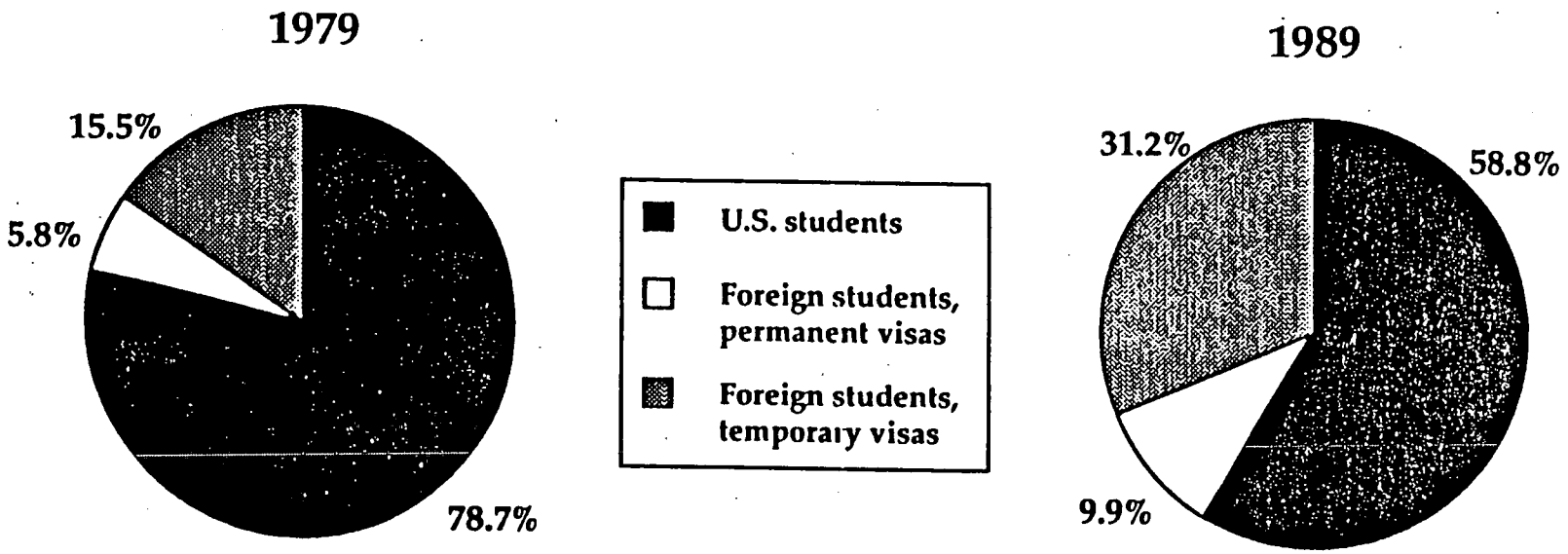


Note: Total Engineering includes: Chemical, Civil, Electrical, Materials Science, and Mechanical Engineering.

Source: March, 1990, report of the Science and Engineering Education Sector Studies Group, National Science Foundation.

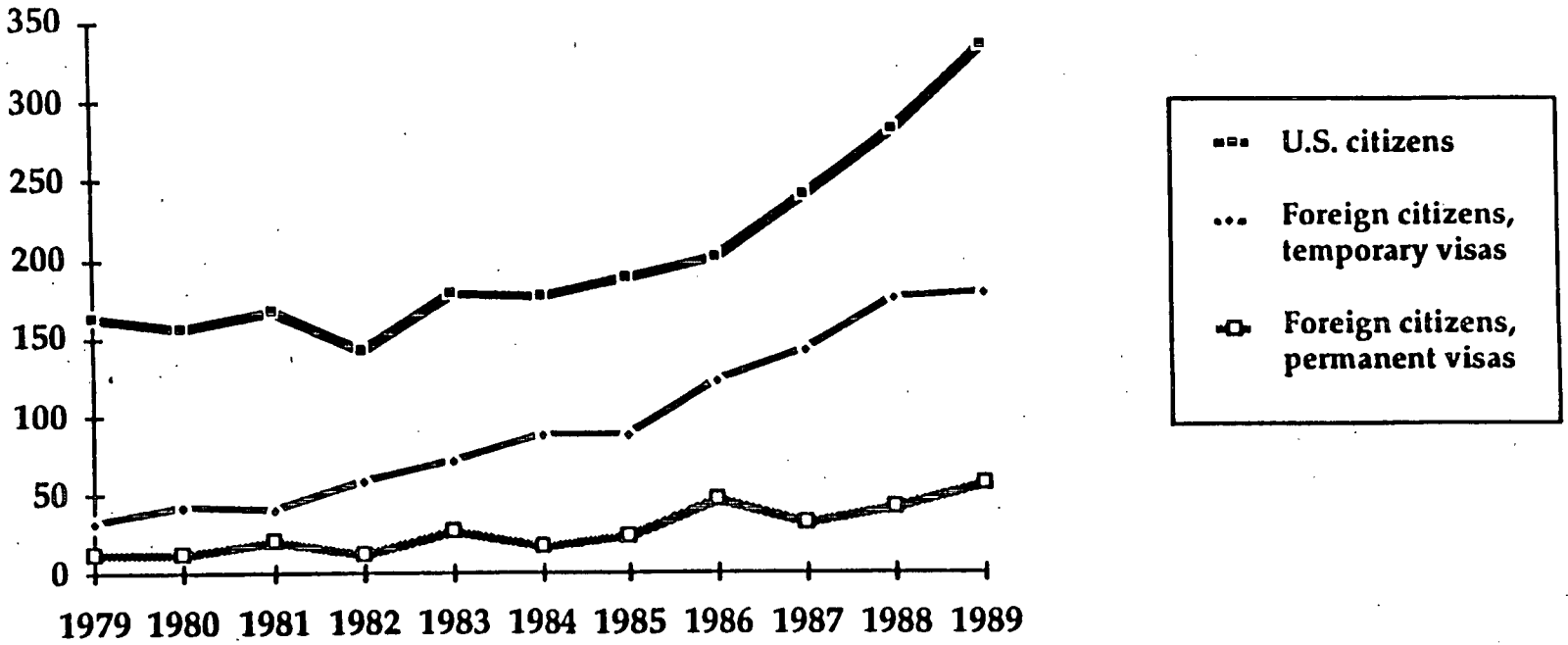
In computer and information science, the decline in U.S. share of doctorates from 1979 to 1989 is even sharper than in science and engineering overall. As shown in Exhibit H-3, foreign students with temporary visas more than doubled their share of doctorates in this period. In terms of absolute numbers, however, the picture is a little brighter. The number of computer/information science doctorates earned by U.S. students approximately doubled from 1979 to 1989, and when foreign students with permanent visas are included, the increase was 125 percent (see Exhibit H-4).

Exhibit H-3: Computer/Information Science Doctorates Received in the U.S., 1979-1989



Source: March, 1990, report of the Science and Engineering Education Sector Studies Group, National Science Foundation.

Exhibit H-4: Computer/Information Science Doctorates Received in the U.S., 1979-1989



Source: March, 1990, report of the Science and Engineering Education Sector Studies Group, National Science Foundation.

That, however, does not directly address the problem of HPC manpower, because as (former NSF Director of Computer/Information Science and Engineering) Gordon Bell noted in a recent paper on "The Future of High Performance Computers in Science and Engineering," the computer science community has not given HPC the attention it deserves. Even more to the point, the vast preponderance of potential users of HPC are trained in other academic disciplines and may be totally ignorant of computational science and its benefits. Although all of the leading American universities either have their own supercomputer facilities or have access to some (see Appendix J), that may not be enough, because as Michael Porter has noted, high quality at the top can mask grave problems elsewhere in the educational system.

In general, the common perception of managers and other experts in various aspects of HPC is that the U.S. is woefully lacking in scientists and engineers with the training and/or experience to design, program, and use supercomputers. This is partly due to the overall shortage of people in science and engineering, but the situation is especially acute, and the implications are especially serious, in High Performance Computing.

HPCC PROGRAM IMPACT

Although the "Basic Research and Human Resources" component of the proposed Federal HPCC Program is the smallest is terms of total funding, it is in many respects the most important component in terms of its bearing upon the ultimate success of the program. Building more and bigger supercomputers will be fruitless unless there is a sufficient number of people who are able and willing to use them.

The goals of the "Human Resources" component of the HPCC Program include:

- **Attaining a level of 1000 computer science Ph.D.s per year by 1995;**
- **Promoting at least 10 interdisciplinary computational science and engineering degree programs;**
- **Upgrading 10 university computer science department to the standards of the current 10 best; and**
- **Upgrading an additional 25 computer science departments to nationally competitive quality.**

Curiously enough, if the number of computer science Ph.D.s continues to grow at the 8 to 9 percent annual rate seen in recent years, more than 1000 doctorates would be produced in 1995 without the HPCC Program. However, it is very uncertain whether this level of growth can be sustained without additional Federal funding, such as that included in the proposed HPCC Program.

As shown in Exhibit H-5, we project that the HPCC Program would increase computer science Ph.D. production (over what would be otherwise attained) by about 19 percent in 1995 and by as much as 35 percent, to more than 2,000, in the year 2000. Masters and bachelors degree output would also be boosted, as suggested in Exhibit H-6, to nearly 11,000 and 23,000, respectively, in 2000 (as compared with about 4,500 and 11,000 in 1990).

Exhibit H-5: Computer/Information Science Doctorates, Scenarios A and B

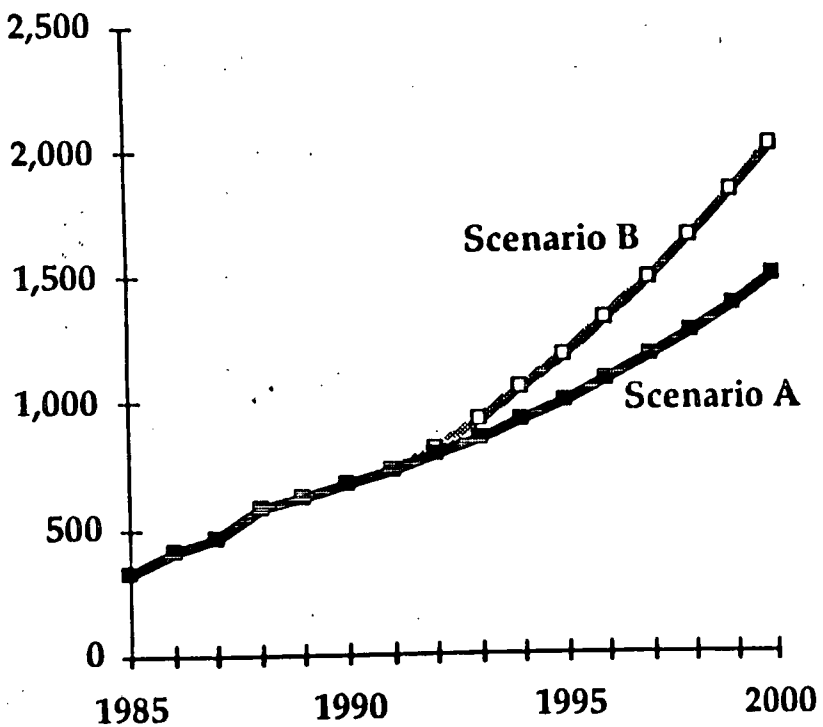
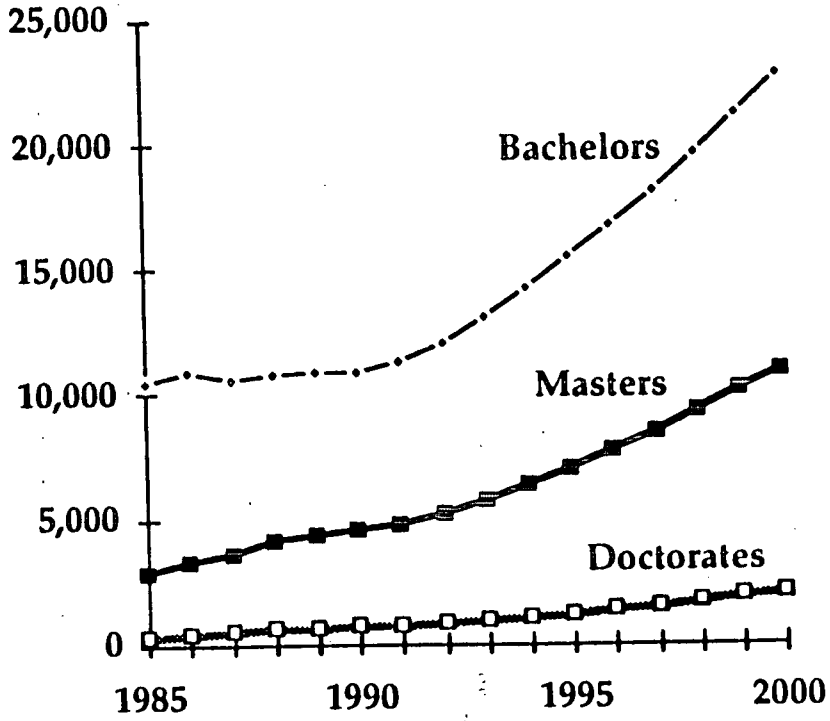


Exhibit H-6: Computer/Information Science Degrees, Scenario B



In the present context, what is more important than the actual number of computer science graduates is the portion of computer scientists who are working in HPC. As noted above, this portion has been rather meager in recent years, but perhaps the prospect of a new source of research funding will attract more computer scientists to HPC. Still more important may be the training of scientists and engineers, other than those already in the computer field, in computational science. Here an old adage may be apt: "It is easier to teach programming to a scientist than to teach science to a programmer."

In light of the perennial paucity of funds in academia, it seems virtually certain that the availability of additional research and education funds for HPC will attract more and better students. Perhaps it is not too much to hope that the HPCC Program will even capture the popular imagination (as MITI's computer initiatives seem to have done in Japan) enough to attract some students away from pursuing MBAs and law degrees.

[This page has been left blank intentionally.]

APPENDIX I - HPC RESEARCH ACTIVITIES

[This page has been left blank intentionally.]

JAPAN

In the past 20 years, Japan has emerged from the shadows to assume a dominant position in the information industry, not only in supercomputers, but across the board. For instance:

- Of the world's top 10 information systems companies (measured by 1989 information systems revenue in **Datamation** magazine's annual survey), the Japanese are ranked 3rd (NEC), 4th (Fujitsu), and 6th (Hitachi);
- In mainframes, Japanese companies are ranked 2nd, 3rd, and 4th;
- In minis, they have the 3rd, 4th, and 6th spots;
- In workstations, they are 4th, 6th, and 9th;
- In PCs, the Japanese are 3rd, 7th and 10th;
- In software, Japanese firms are in 2nd, 4th, and 8th place;
- In peripherals, they hold the 3, 5, 7, and 10 positions;
- In data communications, they are 3rd, 5th, 7th, 8th, and 9th;
- And in vended semiconductors, the Japanese are 1st, 2nd, 3rd, 5th, 7th, and 9th.

In its computer production, Japan has grown at almost twice the rate of the U.S. in recent years -- 26% vs. 14% compound annual growth rate (CAGR), 1977-1987 -- and Japan has become the leading foreign supplier in the U.S. computer market -- 42% of U.S. imports, 16% of U.S. consumption, and $5 billion trade surplus in 1987. (These numbers would be even higher if the shipments of Japanese-owned manufacturing facilities in the U.S. were included.) Japanese firms also compete vigorously against U.S. computer suppliers in every key foreign market; altogether, about one-third of Japan's computer output is exported -- 36% CAGR in export value, 1977-1987.

How did the Japanese do it? Through a carefully coordinated and flexible set of government actions designed to protect and nurture their domestic information industry. As described in Marie Anchordoguy's definitive study, Computers, Inc. (Harvard University Press, 1989), these actions have been of four basic types:

- Protection of domestic markets;
- Financial assistance to key firms;
- Assistance in financing of computer leases (via JECC); and
- Cooperative R&D programs.

The last of these is relevant to the proposed Federal HPCC Program, because it shows the kind of competition which this program is attempting to meet -- and it may even provide a model for the U.S. to consider.

As shown in Exhibit I-1, the Japanese government -- in particular, the Ministry of International Trade and Industry (MITI) -- has sponsored a succession of R&D projects aimed at building the strength of Japanese companies in various aspects of the information industry. To some extent, MITI did "pick winners and losers" in that it selected which technologies would be emphasized at various times and it decided which Japanese companies would be allowed -- or, in some cases, forced -- to participate in each of the projects, and how. (The same could be said, albeit to a lesser extent, of the government-sanctioned and partially government-supported consortia, such as MCC and Sematech, established in the U.S. in the 1980s.) But the "picking winners and losers" characterization distracts attention from an essential point: once the projects were concluded, Japanese companies (whether project participants or not) were quick to apply what had learned in the projects to the development of commercial products, and it was there, and in the subsequent competition in the marketplace, that the real "winners and losers" were ultimately determined.

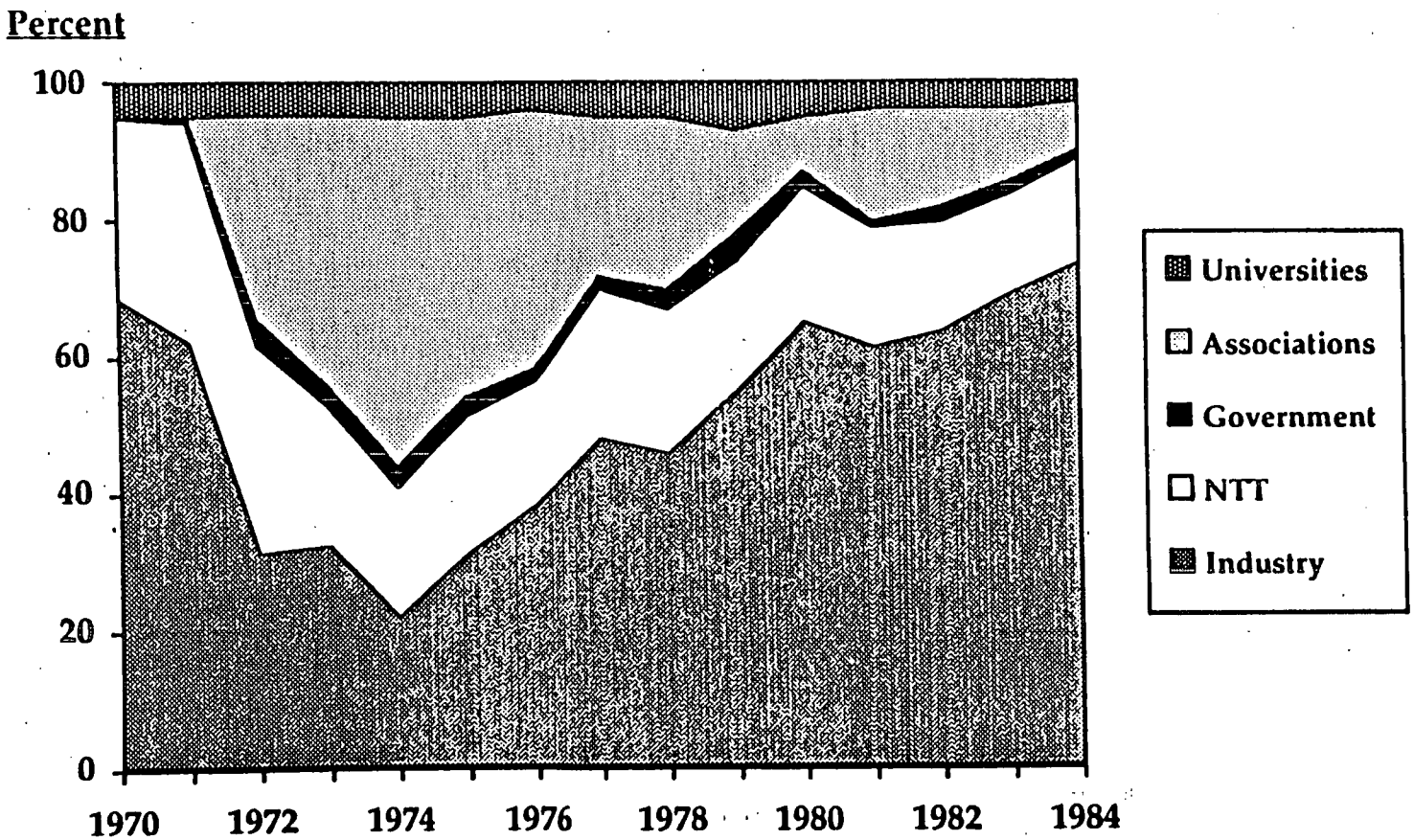
Exhibit I-1: Japanese Computer-Related R&D Projects

When	Project	Funding	1982 Dollars
1962-1965	FONTAC	$2 M	$6.1 M
1966-1971	Super High Performance Computer	$35 M	$90.4 M
1968-1971	DIPS-1	$97 M	$237.3 M
1971-1980	PIPS	$67 M	$109.6 M
1972-1975	3.75 Generation Computers	$235 M	$454.4 M
1973-1975	DIPS-11	$15 M	$27.8 M
1976-1980	VLSI	$323 M	$447.5 M
1976-1981	Software Automation	$30 M	$39.8 M
1979-1986	Opto-electronics	$82 M	$80.4 M
1979-1983	Software Technology	$215 M	$228.8 M
1981-1988	New Function Elements	$20 M	$18.3 M
1981-1989	Supercomputer	$105 M	$94.6 M
1982-1991	Fifth Generation Computer	$357 M	$308.6 M

Sources: Marie Anchordoguy, Computers, Inc., Harvard University Press, 1989; and Gartner Group.

Note: The figures in this exhibit should be interpreted with great caution. Not only are there problems in translating Japanese yen to U.S. dollars (because of wide differences in the exchange rates over the years), but there are even greater difficulties in sorting out Japanese government expenditures from total project budgets (which usually include contributions from the industrial participants). In addition, NTT has sometimes conducted its own parallel "shadow" projects, thereby further intensifying the overall level of R&D (and/or providing a hedge in case the MITI -backed approach did not work out).

Exhibit I-2: Japanese Computer-Related R&D Funding



Source: Kenneth Flamm, Targeting the Computer, Brookings Institution, 1987.

Another important point is that the funding came from a number of sources (see Exhibit I-2). However, because most of the leading universities in Japan are government-run and NTT was, until recently, a government-owned corporation, government funding really consisted of three components. On numerous occasions, there was competition among the various Japanese government agencies in these matters, but when push came to shove, there was general interagency cooperation and concerted action -- which is an important facet of the U.S. HPCC Program, both in its formulation and in its proposed implementation under the aegis of OSTP.

Again we hasten to add that these R&D projects are not solely responsible for Japan's rapid rise to eminence in the worldwide information industry -- but they certainly helped. In the words of a December, 1989, report from the Secretary of Commerce to the House Appropriations Committee on **The Competitive Status of the U.S. Electronics Sector: From Materials to Systems:**

> "Although government support substantially influenced the development of the Japanese computer industry, other factors have had some bearing on the success that the Japanese have enjoyed in the world market. Japan has a culture which promotes cooperation between government, industry and labor to achieve national goals once a consensus is reached. This nation also has a well-educated populace which has provided industry with the skilled human resources it needs and assisted in the creation of a strong technical base. For example, the performance of Japanese high school students on international tests of mathematics and science skills has consistently outpaced those of other major industrialized countries for over a decade....
>
> "Japanese corporations have certain characteristics of their own which have made them formidable competitors. Their managers generally take a long-term view in developing corporate strategy. As a result, they place great emphasis on gaining market share at the expense of short-term profits and [they] price aggressively (even dump) to attain this goal. They have paid close attention to process and production technology. This has allowed the Japanese to become low-cost, high-volume producers. They also understand the importance of linking R&D and manufacturing to ensure efficient technology transfer within the corporation and to build quality and reliability into their products."

What the Japanese have done in the past, however, is not necessarily what they will do in the future. As the times change, they will shift their tactics -- and we in the U.S. should be prepared to shift ours. To quote the Secretary of Commerce's report once again:

> "Government policy in Japan is currently directed toward helping the Japanese computer industry gain technological superiority during the 1990's and ultimately dominate the world market. However, the [Government of Japan's] role in charting a course for the industry has become more complicated, and the problems Japan faces today are somewhat different than they were a decade ago. In fostering research, the government continues to set goals for the national projects and contribute funds for them, but it has found cooperation from Japanese companies harder to obtain. It must contend not only with corporations placing a higher priority on their own internal R&D efforts, but also with the fact that both government and industry must compete for a limited number of qualified researchers in Japan who can work on such exotic technologies as artificial intelligence and superconductivity. In response to this need, the [Government of Japan] has begun to reform its education system. It is pouring more funds into basic research within academia, expanding collaborative efforts between universities and industry, and revising science and engineering curricula to improve Japan's capacity to innovate."

EUROPE

In contrast to Japan, Europe played a major role in the early days of computer development, during and immediately following World War II. However, since the mid-1960s, European computer firms have struggled to keep pace with their American -- and, more recently, their Japanese -- counterparts. This is not to deny, however, the continuing vigor of European scientific research, including that in High Performance Computing. In 1977, when the marketplace in contemporary supercomputers was just being established, the first sale to an institution outside the United States was made to the European Centre for Medium-Range Weather Forecasts (ECMWF) in Reading, England. And in 1979, Cray made additional sales to the British Atomic Weapons Research Establishment in Aldermaston and to the Max Planck Institute for Physics in Garching, West Germany.

The problem for Europe, in HPC as in many other areas, is its political and economic fragmentation. Although it may outnumber and outweigh Japan when taken as a whole, Europe has frequently been the victim of the long-standing rivalries and mistrust among its various countries, and this may be an insurmountable barrier to attaining "critical mass" in R&D projects in rather esoteric areas such as HPC. Nevertheless, like the U.S. -- perhaps even better than the U.S. -- Europe has responded to the Japanese challenge by establishing a number of research efforts pertaining to High Performance Computing (see Exhibit I-3). Although it is dubious that a rival for Cray Research will emerge from this milieu, some smaller contenders have already entered the parallel supercomputer (or minisupercomputer) market with systems based on the Inmos "transputer" chip. Certainly, the entrepreneurial spirit is alive and well in Europe, so we would expect to see additional commercial ventures stemming from European research experience in HPC in the coming years.

Exhibit I-3: Supercomputing Programs in Europe

Program	Funding	Participants	Primary Focus	Comments
European Academic Research Network (EARN)	Information not available at this time.	National Science Foundation Cornell University European Laboratory for Particle Physics; Geneva, Switzerland MCI Communications Corp.	Research and academic computer network; high speed data link to support collaboration/ information exchange both sides of Atlantic.	Founded in 1984 with support of 21 countries; network is similar to U.S. BITNET; recent partnership with NSFnet and EASInet.
Eureka (loose abbreviation for European Research Cooperation Agency)	$200 million earmarked for projects over a five-year period.	Britain France Germany	A broad-based effort to improve Europe's competitiveness in high technology.	Aims at becoming a national center for education and training about supercomputers as well as a source of services for research using supercomputers.

... continued on next page

Exhibit I-3 (cont'd)

Program	Funding	Participants	Primary Focus	Comments
European Computer Industry Research Center (ECRC)	Average annual budget of $7.5 million.	Groupe Bull ICL Siemens	Collaboration in long-range research in artificial intelligence and expert systems.	Established in 1984; the three sponsors share fully in all costs in the laboratory. Sponsoring companies get free licenses to products and systems developed at the laboratory.
European Strategic Program for Research in Information Technology (ESPRIT I &II)	Supermode I: $25 million, 1984-1988; Supermode II: $30 million, 1988-1992	Thom EMI Telmat INMOS RSRE ARSIS Univ. of Grenoble Univ. of Southhampton	Advanced microelectronics Software technology Computer-integrated manufacturing Office automation	ESPRIT has produced a high level of continental scientific cooperation and coordination. It has resulted in technical progress in VLSI and the formation of manufacturing coalitions.

... continued on next page

Exhibit I-3 (cont'd)

Program	Funding	Participants	Primary Focus	Comments
Research in Advanced Communications in Europe (RACE)	$600 million pledged by over 350 organizations.	350 + organizations	Provide Europe with advanced telecommunications service in a timely manner.	Hopes to provide integrated broadband communications (IBC) by 1995.
Information Technology Initiative	$29 million allocated by government.	UK industry and government	Government has been encouraging research and Unix standards.	Since 1979, the UK's trade deficit in information technology and electronics has grown from $440 million to $2.0 billion.
Superconductors Research Programme	$8 million funded by government.	UK industry and government	Research in high Tc superconductivity.	

UNITED STATES

In the words of John Rollwagen, Chairman and CEO of Cray Research: "If it weren't for the U.S. government, there would be no U.S. supercomputer industry." Indeed, there might not even be much of a U.S. computer industry of any kind. The role of government R&D in nurturing innovations which have subsequently been exploited commercially in the computer industry extends all the way back to the wartime work of Eckert and Mauchly, the fathers of Univac. Some more recent examples of commercial high performance computer systems which have roots in government-funded research are given in Exhibit I-4.

Exhibit I-4: U.S. Government-Supported HPC Technologies, Now Commercialized

HPC Technology	University	Company
Multiprocessing, parallelization, vectorization	University of Illinois Rice University	Alliant, Cray, others
Reduced Instruction Set Computing (RISC)	Univ. of Calif. - Berkeley Stanford University	Sun Microsystems MIPS Computer Systems
The Connection Machine	MIT	Thinking Machines Corp.
Very Long Instruction Word processor	Yale University	Multiflow Computer, Inc.
Systolic processors	Carnegie Mellon Univ.	General Electric

The issue at hand, however, is the future. Research spending by the U.S. government has not kept up with inflation in recent years. American spending on non-defense research is about 1.9 percent of gross domestic product (GDP), as compared to 2.8 percent in Japan and 2.6 percent in Germany. This relatively greater investment in non-defense R&D by overseas competitors is doubtless a factor in their success in seizing market share in so many consumer and manufacturing product areas, such as automotive, machine tools, semiconductors, etc. Although U.S. laboratories have provided the research underpinning for many of these products, foreign competitors have been notably more successful in translating this research into commercial sales. Even overall American R&D spending, with defense included, is now lower as a percentage of GDP than in Japan, Germany, and Sweden. For example, according to the National Science Foundation, the United States spent $111.5 billion (2.8 percent of GNP) for R&D in 1988, while the Japanese spent $42.3 billion (2.9 percent of GNP).

To make matters worse, defense R&D spending appears to be facing deep cuts, especially in light of recent developments in Eastern Europe. The conundrum here is that, on the one hand, Pentagon contracting and subsidization of advanced technology research and development has been the primary engine of U.S. technological prowess. On the other hand, while the private sector has assumed a larger role in recent years, private sector R&D investment is now declining, which still leaves a critical role for the Defense Department in developing new technologies. The fact that the Defense Advanced Research Projects Agency (DARPA) has become the principal government funding source for partially civilian technologies -- such as high definition television (HDTV) -- illustrates the nation's habit of looking to the Pentagon for support in generating advanced technology. The dramatic reduction in the Soviet threat and domestic pressures for diverting Defense Department dollars to other uses will radically reduce the funds available to the Pentagon for this purpose in the 1990's. Absent a continuing flow of Federal seed capital for research and development and applied technology, U.S. technology will have even greater difficulty competing against foreign firms, which are already subsidized by their governments. And there is growing danger that America will also become increasingly dependent upon overseas suppliers of critical technologies needed to maintain and develop a modern military establishment.

Although some experts argue that a decline in defense-related R&D spending will have only a minor impact on U.S. technological competitiveness, because rising private sector spending in this area will take up the slack, recent evidence does not bear this out. Industry and other non-federal R&D spending did increase from 36% to 53% of the national total during the last 30 years, but this trend appears to be leveling off. The most recent study by the National Science Foundation indicates that, for the first time in fourteen years, private sector R&D has not kept pace with inflation and that investment in basic research is declining. And another recent survey, by the Center for Innovation Management Studies of Lehigh University, found that U.S. industry is spending less on R&D and reducing its support for research in corporate labs.

Several reasons exist for this downturn. Declining corporate profits, the traditional vulnerability of R&D investment when firms need to trim their operations, the requirement for rapid return on investment, and the threat of hostile takeovers are some of the driving factors. Beyond these lies the sheer cost of developing new technologies, which can extend the resources of the largest companies. But R&D spending by America's international competitors is accelerating and, in some cases, is surpassing the rate of U.S. R&D expenditures. While the U.S. is struggling to stay even with the rate of inflation, Japan is increasing its R&D spending at a rate of 12 percent annually.

What is hidden in all the foregoing figures on relative levels and rates of R&D spending is the allocation of funds among basic research, applied research, and development. In Europe and the U.S., basic research is primarily the province of academia, with some involvement by government laboratories and a few well-heeled large companies. Small companies rarely participate in basic research at all, except possibly through consortia, even though they are regarded (in the U.S. at least) as a primary source of innovation. In applied research, the picture is shifted only slightly, with proportionally greater participation by government and the private sector. Development, however, is regarded as an exclusively private-sector domain, with even consortia being prohibited by antitrust laws (except for *ad hoc* multiple-company teams which occasionally bid on major defense systems).

This is in contrast to the situation in Japan – and in the newly industrialized nations of Southeast Asia -- where the emphasis is upon development first, applied research second, and basic research hardly at all. And unlike the U.S., there may be cooperation in development in Japan -- although this seems to have little moderating effect upon the overall (very intense) competition. (One of the first actions in Japan's industrial policy for computers was to pass a measure – in 1957 -- exempting the computer industry from the antitrust laws imposed by the U.S. Occupation and even allowing MITI to force companies to participate in cartels when the government felt it necessary.) But while the amount of Japanese spending on basic research has been quite small, as compared with that in Europe and the U.S., that is now changing. Beginning with the establishment of the Fifth Generation Computer Project in 1981, Japan has been increasing its emphasis on basic research, not only in academia and government, but in the private sector as well. Perhaps this is an indication that Japan feels that it has attained sufficient economic status that it can now afford this "luxury" (which it certainly could not afford during post-war reconstruction) and/or that it can no longer depend upon the U.S. (and Europe) to sustain the level of basic research it needs to fuel its future growth. At any rate, it appears that U.S. pre-eminence in basic science is also going to be challenged in the years ahead.

These trends seem to indicate that the United States is headed for second-rate economic status unless there is a turnaround in attitudes and investment practices. Without the defense spending dynamo to drive it, America's technological engine does not seem to be capable of generating and maintaining the momentum necessary to keep pace with Japan. Although Europe is running a distant third in the HPC race at present, the renewed vitality there, resulting from the impending 1992 unification and the breakup of the Eastern bloc, might even put it in a position to pass the U.S. by the end of the decade. Certainly, the current European renaissance is evidence of a willingness to take decisive action in response to the Japanese challenge. It is time for the U.S. to act decisively, too.

[This page has been left blank intentionally.]

APPENDIX J - SUPERCOMPUTING FACILITIES

[This page has been left blank intentionally.]

Exhibit J-1: Supercomputing Facilities at U.S. Universities

Institution	Supercomputer	Institution	Supercomputer
ALABAMA		CALIFORNIA, continued	
Alabama Supercomputer Authority	Cray X-MP/24	California State University, Sacramento	Elxsi 6400; Multiflow 14/300
University of Alabama	IBM 3090-400E/VF	Humboldt State University	Sequent S-81
ARIZONA		John F. Kennedy University	Encore Multimax 310
Arizona State University	Convex C2; Cray X-MP/14se; IBM 3090-500E/3VF	Point Loma Nazarene College	Encore Multimax 520
Northern Arizona University	IBM ES/9000-210/VF	San Jose State University	Sequent 8000
University of Arizona	Convex C120, C240; IBM 3090-300E/VF; SCS 40	Stanford University	IBM 3090-600E/6VF
CALIFORNIA		University of California, Berkeley	Ardent Titan; Cray X-MP/14; IBM 3090-200S/VF, 3090-300E/2VF
California Institute of Technology (JPL)	Cray X-MP/18	University of California, Davis	DEC VAX 6000-410 VF
California Polytechnic University	Pyramid 90X; IBM 3090-400E/VF	University of California, Los Angeles	IBM 3090-600J/6VF; SCS 40
California State University, Hayward	Pyramid 9805; Sequent S-27	University of California, San Diego	(See Exhibit J-2)
		University of San Diego	Pyramid 9805
		University of Southern California	Alliant FX/180, FX/2800; IBM 3090-180E/VF

Sources: Alliantr Computer Systems; CDC; Cray Research; DEC; IBM; Thinking Machines; Sidney Fernbach; Supercomputing Review; Sidney Karin and Norris Parker Smith, The Supercomputer Era, Harcourt Brace Jovanovich, 1987; Charles H. Warlick, ed. Directory of Computing Facilities in Higher Education, University of Texas at Austin, 1990.

... continued on next page

Exhibit J-1 (cont'd)

Institution	Supercomputer
COLORADO	
Colorado State University	CDC Cyber 205/422
University of Colorado, Boulder	Alliant FX8; Sequent 21000
University of Colorado, Denver	Intel IPSC; Pyramid 90X; Sequent Symmetry
CONNECTICUT	
University of Connecticut	IBM 3090-150E/VF, ES/9000-580/3VF
Yale University	IBM 3090-180E/VF
DISTRICT OF COLUMBIA	
Howard University	Alliant FX/2800; IBM 3090-180J/VF
DELAWARE	
University of Delaware	IBM 3090- 300E/3VF
FLORIDA	
Florida Institute of Technology	Ardent
FLORIDA, continued	
Florida State University	Cray Y-MP/432; ETA 10Q-264; Thinking Machines CM-2/64K
University of Florida	IBM 3090-600E/6VF
University of North Florida	Sequent B8000
University of South Florida	IBM 3090-300E/VF
GEORGIA	
Georgia Institute of Technology	CDC Cyber 180/990; Pyramid 90X; Sequent S-81
University of Georgia	CDC Cyber 205/622; IBM 3090-400E/2VF
HAWAII	
University of Hawaii, Manoa	Alliant FX8; IBM 3090-200E/VF
ILLINOIS	
Illinois Benedictine College	Sequent Balance
Illinois Institute of Technology	Encore Multimax

... continued on next page

Exhibit J-1 (cont'd)

Institution	Supercomputer
ILLINOIS, continued	
North Central College	Sequent S-27
Northern Illinois University	Encore Multimax 6
Northwestern University	Pyramid
Southern Illinois University	IBM 3090-150E/VF
University of Illinois, Urbana-Champaign	Alliant FX/2800; Convex C220; Cray X-MP/48; two DEC VAX 6000-410 VFs; IBM 3090-300J/3VF; Pyramid 90X; Sequent B-8000; (See also Exhibit J-2)
William Harper Rainey College	Sequent
INDIANA	
Indiana State University	Sequent
Indiana University	IBM 3090-120E/VF
Indiana Univ.-Purdue Univ./Indianapolis	IBM 3090-180J/VF
Purdue University	CDC Cyber 205/422; ETA-10P; IBM 3090-180E/VF; Sequent Symmetry, S-37
IOWA	
Iowa State University	SCS 40
University of Notre Dame	Convex C120
University of Iowa	Alliant FX8; Encore Multimax; IBM 3090-200E/VF
KANSAS	
Bethany College	Encore
Sterling College	Encore
KENTUCKY	
University of Kentucky	IBM 3090-600J/6VF
University of Louisville	IBM 3090-400E/VF
LOUISIANA	
Louisiana State University, Baton Rouge	FPS 264, 500; IBM 3090-600E/2VF
University of Southwestern Louisiana	IBM 3090-200/VF; Pyramid 90X

... continued on next page

Exhibit J-1 (cont'd)

Institution	Supercomputer
MAINE	
University of Maine	IBM 3090-180E/VF
MASSACHUSETTS	
Boston University	Encore Multimax; IBM 3090-200/2VF
Framingham State College	Encore Multimax 320
Massachusetts Institute of Technology	Cray-2/4-256
Northeastern University	Pyramid 98X
Tufts University	Encore Multimax 310
MICHIGAN	
Michigan State University	BBN GP-1000; Convex 220; IBM 3090-180/VF
Michigan Technological University	Alliant FX/2800s (two); Encore Multimax 520; Sequent 8000
University of Michigan	IBM 3090-600E/2VF
MINNESOTA	
Bethany College and Seminary	Pyramid 90X
Hamline University	Sequent S-27
University of Minnesota	Cray X-MP/48, Cray-2S/4-128; Encore Multimax 520; dual IBM 3090-600J/6VFs; Intel Hypercube; Thinking Machines CM-2/32K
University of Minnesota, Duluth	Encore Multimax
MISSISSIPPI	
University of Mississippi	CDC Cyber 205; ETA-10Q
MISSOURI	
University of Missouri	IBM 3090-170J/VF
NEBRASKA	
University of Nebraska, Omaha	Sequent B8000

... continued on next page

Exhibit J-1 (cont'd)

Institution	Supercomputer
NEW HAMPSHIRE	
Plymouth State College/UNH	Pyramid 98X
NEW JERSEY	
Rutgers, The State University	Pyramid 9810
NEW MEXICO	
New Mexico State University	FPS T-100; Sequent S-26,
University of New Mexico	Sequent S-27
NEW YORK	
Clarkson University	Alliant FX8
Cornell University	IBM 3090-200J/2VF; (See also Exhibit J-2)
CUNY, University Computing Center	IBM 3090-400E/2VF
King's College	Encore Multimax 310
New York University	Elxsi 6400
Rensselaer Polytechnic Institute	IBM 3090-200S/2VF; Sequent 2100
NEW YORK, continued	
SUNY/Binghampton	IBM 3090-180J/VF
SUNY/Buffalo	Encore Multimax; IBM 3090-180E/VF; Intel Hypercube
Syracuse University	Alliant FX80; Encore Multimax 520/16, 320/20; Thinking Machines CM-1, CM-2
NORTH CAROLINA	
North Carolina State University	IBM 3090-180J/VF
North Carolina Supercomputing Center	Cray Y-MP8/432
University of North Carolina, Chapel Hill	Convex C240; IBM 3090-170J/VF
University of North Carolina, Wilmington	Sequent B8
Wake Forest University	Convex C120
NORTH DAKOTA	
North Dakota State University	IBM 3090-200E/VF

... continued on next page

Exhibit J-1 (cont'd)

Institution	Supercomputer
OHIO	
Air Force Institute of Technology	Elxsi 6420; Encore Multimax 320
Akron University	IBM 3090-200/VF
Ohio Supercomputer Center	Convex C1; Cray Y-MP8/864
OKLAHOMA	
Oklahoma State University	IBM 3090-200S/VF
University of Oklahoma	Encore Multimax
OREGON	
Oregon Institute of Technology	Sequent S-27
Portland State University	Sequent S-27
Oregon State University	FPS M64; Sequent 21000
University of Oregon	Convex C1
PENNSYLVANIA	
Drexel University	IBM 3090-150E/VF
PENNSYLVANIA, continued	
Lehigh University	Ardent T1-253
Pennsylvania State University	IBM 3090-600S/6VF; ES/9000-320/VF
Pittsburgh Supercomputer Center	(See Exhibit J-2)
Temple University	FPS 264
University of Pennsylvania	Ardent Titan II; IBM 3090-200E/2VF
Villanova University	Pyramid
RHODE ISLAND	
Brown University	IBM 3090-180E/VF; IBM ES/9000-320/VF
SOUTH CAROLINA	
Clemson University	DEC VAX 6000-410 VF
University of South Carolina	DEC VAX 6000-440 VF
SOUTH DAKOTA	
Augustana College	Encore Multimax 310

... continued on next page

Exhibit J-1 (cont'd)

Institution	Supercomputer
TENNESSEE	
University of Tennessee	IBM 3090-200E/2VF
TEXAS	
Houston Area Research Center	NEC SX-2
Southwestern University	Sequent B8000
St. Mary's University	Alliant FX4
Texas A&M University	Cray Y-MP2/216; IBM 3090-200E/VF
Texas Tech University	Ardent Titan
University of Texas, Arlington	Alliant FX80, Convex C220
University of Texas, Austin	Cray X-MP/14se, X-MP/24; Encore Multimax; IBM ES/9000-720/2VF
University of Texas, Dallas	Convex C1; Encore Multimax
Univ. of Texas Anderson Cancer Center	Alliant FX140
University of Texas System	Ardent Titan; Convex C1; Cray X-MP/14, X-MP/24
UTAH	
University of Utah	IBM 3090-600S/6VF
VIRGINIA	
Old Dominion University	IBM 3090-180/VF
University of Virginia	Convex C1; IBM 3090-150E/VF
Virginia Polytechnic Institute	Convex C210; IBM 3090-300E/3VF; Sequent 8000
WASHINGTON	
Pacific Lutheran University	Intel Hypercube
Seattle University	Encore Multimax 310
University of Washington	Convex C210; IBM 3090-300E/3VF; Sequent S-81
Washington State University	Sequent B8
Western Washington University	Sequent S-81
WEST VIRGINIA	
West Virginia University	IBM 3090-300E/3VF

Exhibit J-2: Supercomputer Centers Sponsored by the National Science Foundation

Institution	Funding	Supercomputer	Primary Focus	Comments
Cornell National Supercomputer Facility, Cornell University	The NSF allocates 60% of total computing time available at each of the five NSF-supported centers.	Two IBM 3090-600Js (6 vector processors each; potential for 12-way parallelism; total peak throughput > 1.5 Gflops). Broad range of visualization and animation equipment.	Parallel computing as well as vector and scalar computing for all research disciplines in academia and industry nationwide.	Strategic users receive dedicated assistance with parallelization, visualization; additional allocations. Each application may use up to 1 Gbyte virtual memory. Smart Node program provides on-site training and consulting at more than 50 remote institutions.
National Center for Supercomputing Applications (NCSA), University of Illinois, Urbana/Champaign	Same as above. Industrial partners each contribute up to $1 million/year.	Cray-2S/4-128; Cray X-MP/48, 128 Mword SSD; 32K processor Connection Machine; Alliant FX80; Convex C240; Amdahl 5860 back-end; VAX 785 front-ends.	Interdisciplinary research; intent is to create a community of resident and visiting scholars who will explore new ways to apply the computational powers of high performance computing systems to the sciences and arts.	NCSA concentrates upon improving imaging, algorithms, and software tools; center has acquired a large number of workstations and personal computers (*e.g.*, Sun, Macintosh, IBM PC) for use by resident and visiting researchers.

Sources: **Supercomputing Review; Sidney Karin and Norris Parker Smith, The Supercomputer Era, Harcourt Brace Jovanovich, 1987.**

... continued on next page

Exhibit J-2 (cont'd)

Institution	Funding	Supercomputer	Primary Focus	Comments
Pittsburgh Supercomputing Center (PSC)	NSF funding same as above.	Cray Y-MP8/832, UNICOS; 32K processor Connection Machine	Basic research; provides quality supercomputer capability to the scientific and engineering communities.	The PSC is a joint effort of Carnegie-Mellon University and the University of Pittsburgh, together with Westinghouse Electric Corp.
John von Neumann Center for Scientific Computing (JVNC)	NSF; also received $12 million construction grant from the State of New Jersey.	CDC Cyber 205; ETA-10;VAX 8600 cluster served as front-end.	The JVNC was to provide state-of-the-art computing and communications to university, government and industrial researchers.	Located in Princeton, NJ, and operated by Consortium for Scientific Computing (mostly northeastern universities). Following CDC's withdrawal from the supercomputer business, the NSF did not renew funding for the JVNC. It became inactive as of April 30, 1990. The actual network will stay in place.
San Diego Supercomputer Center (SDSC), University of California, San Diego	NSF funding same as Cornell, NCSA, and PSC.	Cray X-MP/48; Cray Y-MP8/864; Alliant FX/2800; Supertek S-1 minisuper.	Scientific research: biochemistry, physics, mechanical and electrical engineering, computational fluid dynamics.	Affiliate relationships with SDSC have been established by Aerojet General, Amoco, Battelle Memorial Institute, International Telephone and Telegraph (ITT), MACOM Linkabit, Omnibus, Science Applications International (SAIC), and the Rohr Corp.

Exhibit J-3: Supercomputing Facilities at Federal Research Institutions

Institution	Funding	Supercomputer	Primary Focus	Comments
U.S. Army Ballistic Research Laboratory (BRL)	Available upon request to qualified DOD and government agencies and their contractors.	Cray-2, 256 Mwords, UNICOS; Cray X-MP/48, 128 Mword SSD, UNICOS.	Supercomputing access for defense department associates and contractors.	Provides both unclassified and classified facilities through on-site operations, dial-up, or networks.
National Cancer Institute (NCI)	Department of Health and Human Services	Cray X-MP/28.	Biomedical research.	NCI's Advanced Scientific Computing Laboratory is at the Frederick Cancer Research Facility, Frederick, MD.
National Center for Atmospheric Research (NCAR)	Operated by a non-profit consortium of universities; primarily supported by NSF.	Cray X-MP/48, 256 Mword SSD, COS; Cray X-MP/18, UNICOS; 8K processor Connection Machine; IBM 4381 front-end.	Atmospheric, oceanographic, and related sciences.	About 40% of available resources are reserved for researchers on the NCAR staff, 40% goes to scientists outside NCAR; and the remainder is used for joint projects.
National Energy Research Supercomputer Center (NERSC)	Department of Energy	Cray-1/S; Cray X-MP/22; three Cray-2s (one is 8-processor).	High energy physics, materials sciences, chemical sciences, heavy ion fusion, health and environmental research, applied plasma physics.	One of the pioneers in supercomputer operations and networking; formerly called the National Magnetic Fusion Energy Computer Center.

Sources: Supercomputing Review; Sidney Karin and Norris Parker Smith, The Supercomputer Era, Harcourt Brace Jovanovich, 1987.

... continued on next page

Exhibit J-3 (cont'd)

Institution	Funding	Supercomputer	Primary Focus	Comments
National Institute of Standards & Technology (NIST)	Department of Commerce	CDC Cyber 205; CDC Cyber 180/846 front-end.	Research in thermal physics, fire modeling, structural collapse, and material science problems.	Many cooperative and collaborative projects with scientists and engineers from universities and industry.
NASA Supercomputing Facilities	Office of Space Sciences and Applications; remote users spend about 30 percent on supercomputers.	Two CDC Cyber 205s, two Cray X-MPs,two Cray-2s, at least four Cray Y-MPs, and a number of parallel systems from Intel, MasPar, Thinking Machines, etc. plus a one-of-a-kind Massively Parallel Processor made by Goodyear Aerospace Corporation.	Aerospace research, satellite image processing.	Research centers with supercomputers include Ames, Goddard, Langley, Lewis, and Marshall.

Exhibit J-4: Academic Affiliates of Supercomputer Centers

Key:

CTC: Cornell Theory Center-Cornell National Supercomputer Facility (consortium members)

HARC: Houston Area Research Center (Consortium members)

NCSA: National Center for Supercomputing Applications (Illinois affiliates)

PSC: Pittsburgh Center (academic affiliates)

SCC: Single-campus computer center (may be networked)

SDSC: San Diego Supercomputer Center (consortium members)

NSF: Site of NSF-supported center

NCAR: Member of National Center for Atmospheric Research consortium

Note: Applicants not affiliated with consortium members have equal access to NSF-allocated supercomputer time at NSF-supported centers

Source: Sidney Karin and Norris Parker Smith, The Supercomputer Era, Harcourt Brace Jovanovich, 1987.

... continued on next page

Exhibit J-4 (cont'd)

Northeast and Middle Atlantic States:

Brown University (PSC)
Carnegie-Mellon University (PSC, NSF)
Case Western University (PSC)
Columbia University (PSC)
Cornell University (NSF, NCAR)
Cornell Medical Center (CTC)
University of Delaware (CTC)
Drexel University (NCAR)
George Washington University (NCSA)
Harvard University (NCSA, PSC, NCAR)
John Hopkins University (CTC, PSC, NCAR)
Institute for Advanced Study (PSC)
Lehigh University (PSC)
University of Maryland (SDSC, PSC, NCAR)
Massachusetts Institute of Technology (NCAR)
State University of New York at Albany (NCAR)
State University of New York at Binghamton (CTC)
State University of New York at Stony Brook (PSC)
University of Pennsylvania (PSC)
Pennsylvania State University (CTC, NCSA, PSC, NCAR)
University of Pittsburgh (PSC, NSF)
Princeton University (NCAR)
Rensselaer Polytechnic Institute (CTC)
University of Rhode Island (NCAR)
University of Rochester (CTC)
Rockefeller University (CTC)
Rutgers University (PSC)
Syracuse University (CTC)
Temple University (PSC)
Woods Hole Oceanographic Institution (NCAR)
Yale University (PSC, NCAR)

... continued on next page

Exhibit J-4 (cont'd)

Far Western States:

University of Alaska (NCAR)
Agouron Institute (SDSC)
University of Arizona (NCAR)
University of California, Berkeley (SDSC, SCC)
University of California, Davis (SDSC, NCAR)
University of California , Irvine (SDSC)
University of California, Los Angeles (CTC, SDSC, NCAR, SCC)
University of California, Riverside (NCSA, SDSC)
University of California, San Diego (NSF, SDSC)
University of California, San Francisco (SDSC)
University of California, Santa Barbara (SDSC)
University of California, Santa Cruz (SDSC)
California Institute of Technology (SDSC, NCAR)
University of Hawaii, (SDSC, NCAR)
Naval Postgraduate School (NCAR)
National Optical Astronomy Observatories (SDSC)
University of Nevada (NCAR)
Oregon State University (NCSA, NCAR)
Salk Institute (SDSC)
San Diego State University (SDSC)
Research Institute of Scripps Clinic (SDSC)
Scripps Institution of Oceanography (U.C. San Diego) (SDSC, NCAR)
University of Southern California (SDSC)
Stanford University (NCSA, SDSC, NCAR)
University of Washington (NCSA, SDSC, NCAR)

Canada:

University of Calgary (SCC)
McGill University (NCAR)
Ontario University (shares with Toronto)
University of Toronto (SCC, NCAR)